Functional and evolutionary insights from the genomes of three parasitoid Nasonia species

Werren, John H.; Richards, Stephen; Desjardins, Christopher A.; Niehuis, Oliver; Gadau, Juergen; Colbourne, John K.; Beukeboom, Leo W.; Desplan, Claude; Elsik, Christine G.; Grimmelikhuijzen, Cornelis J. P.

Published in:
Science

DOI:
10.1126/science.1178028

Publication date:
2010

Supporting Online Material for

Functional and Evolutionary Insights from the Genomes of Three Parasitoid *Nasonia* Species

The *Nasonia* Genome Working Group†

†To whom correspondence should be addressed. E-mail: werr@mail.rochester.edu (J.H.W.); stephenr@bcm.tmc.edu (S.R.)

Published 15 January 2010, *Science* **327**, 343 (2010)
DOI: 10.1126/science.1178028

This PDF file includes:

Other Supporting Online Material for this manuscript includes the following: (available at www.sciencemag.org/cgi/content/full/327/5963/343/DC1)

Tables S1, S5, S6, S17, S24, S43, S50, and S55

SUPPORTING ONLINE MATERIAL

SUPPORTING TEXT

Contributions and Acknowledgements

List of Contributors

Project Coordination and Major Contributors: John H Werren, Stephen Richards, Christopher A Desjardins, Oliver Niehuis, Jürgen Gadau, John K Colbourne.

Sequencing & Assembly PI: Richard A Gibbs.

Major Analysis Category Coordinators: Leo W Beukeboom, John K Colbourne, Claude Desplan, Christine G Elsik, Jürgen Gadau, Cornelis J P Grimmelikhuijzen, Paul Kitts, Jeremy A Lynch, Stephen Richards, David M Shuker, Christopher D Smith, John H Werren, Evgeny M Zdobnov.

Nasonia White Paper Authors: Leo W Beukeboom, Claude Desplan, Jürgen Gadau, Jeremy A Lynch, Stephen Richards, David Rivers, John H Werren, Louis van de Zande.

Sequencing: Sequencing Management: Richard A Gibbs, Donna M Muzny, **Library Production:** Dean Chavez, Rachel Edwards, Kashif Hirani, Angela J Johnson, Sandra L Lee, Lynne V Nazareth†, Ling-Ling Pu, Stephen Richards, Selina Vattathil, **Sanger Sequencing:** Joseph Chacko, Mimi N Chandrabose, Andrew G Cree, Marvin D Dao, Huyen H Dinh, Ramatu A Gabisi, Sandra Hines, Jennifer Hume, Shalini N Jhangian, Vandita Joshi, Christie L Kovar†, Lora R Lewis, Yih-shin Liu, John Lopez, Margaret B Morgan, Ngoc B Nguyen, Geoffrey O Okwuonu, San J Ruiz, Jireh Santibanez, Rita A Wright, **Sequence Production Informatics:** Charles Moen, Ryan J Lozado, Gerald R Fowler†, **Genome Assembly:** Huaiyang Jiang, Stephen Richards, Kim C Worley†, **Illumina Sequencing and Alignment:** Lesley Chaboub, Yi Han†, Huaiyang Jiang, Jeffrey G Reid, Stephen Richards, **EST and BAC library production and analysis:** Jade Carter, Jeong-Hyeon Choi, John K Colbourne†, Phat M Dang, Christopher A Desjardins, Rachel Edwards, Donald G Gilbert, Zhao Lai, Monica C Munoz-Torres, Deodoro C S G Oliveira, Phillip San Miguel, Zachary Smith, Jeanne Romero-Severson, Wayne B Hunter.

Mapping & Genetics: Victor H Anaya, Leo W Beukeboom, Charles Claudianos, Alexandre S Cristino, Christopher A Desjardins, Rachel Edwards, Jürgen Gadau, Joshua D Gibson , David Loehlin, Albert Kamping, Tosca Koevoets, Oliver Niehuis, John G Oakeshott, Bart A Pannebakker, David M Shuker, Louis van de Zande, John H Werren, **MicroArrays:** Jeong-Hyeon Choi, John K Colbourne†, Christopher A Desjardins, Rachel Edwards, Joshua D Gibson, Bobak Kechavarzi, Heewook Lee, Jacqueline A Lopez, Anoop Mayampurath, Oliver Niehuis, Vikas R Pejaver, Andreas Rechtsteiner, **GC content and Methylation:** Navin Elango, Eran Elhaik, Christine G Elsik, Dan Graur, Tatsuhiko Kadowaki, Ryszard Maleszka, Zuogang Peng, Justin T Reese, Megan Riddle,

Carol Trent, Soojin V Yi, **Microsatellites:** Bart A Pannebakker, David M Shuker, **Heterchromatin and repeats:** Rochelle A Clinton, Thomas H Eickbush, Henry C Hunter IV, Jay Kim, Marcella A McClure, Christopher D Smith†, Deborah E Stage, **Telomeres:** Hugh M Robertson **Micro RNAs and tRNAs:** Juan M Anzola, Susanta K Behura, Christine G Elsik, Daniel Gerlach, Darren E Hagen, Monica C Munoz-Torres, Evgeny M Zdobnov, **Cis-regulatory sequences:** Jaebum Kim, Alexandre V Morozov, Saurabh Sinha.

Automated Annotation: Hsiu-Chuan Chen, Christopher P Childers, Christine G Elsik, Olga Ermolaeva, Donald G Gilbert, Darren E Hagen, Wratko Hlavina, Yuri Kapustin, Boris Kiryutin, Paul Kitts, Terence Murphy, Donna Maglott, Stephen C Pratt, Kim Pruitt, Justin T Reese, Stephen Richards, Victor Sapojnikov, Victor Solovyev, Alexandre Souvorov, Mario Stanke, Kim C Worley, Stefan Wyder, Evgeny M Zdobnov, Lan Zhang, **Phylogenomic analysis:** Thomas Junier, Evgenia V Kriventseva , James B Whitfield, Stefan Wyder, Evgeny M Zdobnov† **KEGG pathway analysis:** Peer Bork†, Jean Muller, Takuji Yamada, **Protein domain analysis:** Erich Bornberg-Bauer†, Andrew D Moore, Andreas Schüler, Arndt Telschow.

Manual Annotation and Gene Analysis: Cuticle proteins: Robert S Cornman, Judith H Willis, **Development:** Claude Desplan, Jeremy A Lynch, Monica C Munoz-Torres, Miriam Rosenberg, **Diapause proteomics:** Florian Wolschin, **Environmental Response Genes:** May R Berenbaum, Charles Claudianos, Alexandre S Cristino, Reed M Johnson, John G Oakeshott, Hilary Ranson, **Hexamerines:** Angel R Barchuk, Márcia M G Bitondi, Alexandre S Cristino†, Francis M F Nunes, Zilá L P Simões, **Immunity:** Seth R Bordenstein†, Ching Crozier, Ross Crozier, Jay D Evans†, Dan Hultmark, Timothy B Sackton, Helge Schlüns, Robert M Waterhouse, Michael Williams, Evgeny M Zdobnov, Courtney N Zecher, Zou Zhen, **Informatics:** Christopher P Childers, Christopher A Desjardins, Christine G Elsik†, Darren E Hagen, Justin T Reese, Stephen Richards, Lan Zhang, **Ion Channels:** Agata N Bera, Stefan Gründer, Andrew K Jones, Kristin Lees, David B Sattelle, Andreas Springauf, **Meiosis:** John M Logsdon, Jr, Danielle J Mazur, Andrew M Schurko, **Neuropeptides:** Rinaldo C Bertossa, Giuseppe Cazzamali, Cornelis J P Grimmelikhuijzen†, Frank Hauser, Susanne Neupert, Reinhard Predel, Martina Schneider, Elisabeth Stafflinger, Yoshiaki Tanaka, Michael Williamson, **Odorant binding proteins and chemoreceptors:** Sylvain Foret, Hugh M Robertson, Kevin W Wanner, **Oxidative phosphorylation:** Jürgen Gadau, Joshua D Gibson, Oliver Niehuis, Deodoro C S G Oliveira, Stephen C Pratt, John H Werren, **Phermones:** Oliver Niehuis, Joachim Ruther, Thomas Schmitt, Jürgen Gadau, **RNAi:** Jeremy A Lynch, Deodoro C S G Oliveira, **Royal Jelly Proteins:** Stefan Albert, Robert Kucharski, Ryszard Maleszka†, **Sex determination:** Rinaldo C Bertossa, Leo W Beukeboom, Albert Kamping, Bart A Pannebakker, Deodoro C S G Oliveira, Louis van de Zande, Eveline C Verhulst, **Venoms:** Maarten Aerts, Marleen Brunain, Dirk C de Graaf†, Christopher A Desjardins, Bart Devreese, Frans Jacobs.

Comparative genomics of *Nasonia* Species: Victor H Anaya, Rhitoban RayChoudhury, Christopher A Desjardins, Rachel Edwards, Oliver Niehuis, Deodoro C S G Oliveira, John H Werren†, James B Whitfield, **Cytonuclear-genic incompatibilities:** Christopher A Desjardins, Rachel Edwards, Joshua D Gibson, Oliver Niehuis, Jürgen Gadau, John H. Werren, **Symbionts and Lateral gene transfer:** Erich Bornberg-Bauer†, Michael E. Clark, Alistair C Darby, Gregory D D Hurst, Andrew D Moore, Deodoro C S G Oliveira, Andreas Schüler, Arndt Telschow, Timothy E Wilkes.

† = group leader

Acknowledgments

We thank the following people for their assistance: Joseph Anderson, Amanda Avery, Francisco Camara, Jade Carter, Mimi Chandrabose, Hsiu-Chuan Chen, Robert Cornman, Ching Crozier, Ross Crozier, Marvin Dao, Thomas Eickbush, Mary Fredendall, Kathy Giardina, Jonathan Giebel, Roderic Guigo, Kevin Hackett, Kathryn Haley, Sandra Hines, Kashif Hirani, Jennifer Hume, Wayne Hunter, Frans Jacobs, Angela Johnson, Vandita Joshi, Zhao Lai, Cindy Landry, Yih-shin Liu, John Lopez, Charles Moen, Alexandre Morozov, Daven Presgraves, Kim Pruitt, Carson Qu, Gabisi Ramatu, San Ruiz, Joachim Ruther, Phillip San Miguel, Jireh Santibanez, Victor Sapojnikov, Justin Sysol, Michael Williams, Hlavina Wratko, Rita Wright, and Zhen Zou. The sequencing of the *Nasonia* genomes was funded by the National Human Genome Research Institute (NHGRI U54 HG003273). Work in the Werren laboratory was funded by grants from the National Science Foundation (FIBR-0328363), National Institutes of Health (5R01GM070026-04 and 5R24GM084917-02), and Indiana 21st Century Research and Technology Fund. Work at the Center for Genomics and Bioinformatics was supported in by the Indiana METACyt Initiative of Indiana University, Lilly Endowment, Inc, and Indiana 21st Century Research and Technology Fund. Work at NCBI was supported by the Intramural Research Program of the NIH, National Library of Medicine. Oliver Niehuis and Florian Wolschin acknowledge the Alexander von Humboldt Foundation for Feodor Lynen Research Fellowships for Postdoctoral Researchers. Work in the Beukeboom lab was funded by grants of the Netherlands Organization for Scientific Research (Pioneer grant ALW 833.02.003 to LWB and Veni grant 863.08.08 to BAP). Work at NYU was supported by grant NIH R01GM064864 to Claude Desplan. Work in the Elsik lab was supported by a grant from the United States Department of Agriculture National Research Initiative (USDA NRI 2008-35302-18804). Work in the Grimmelikhuijzen lab was supported by the Danish Research Agency (Research Council for Nature and Universe), and Novo Nordisk Foundation. Work by Christopher Smith was supported by the Drosophila Heterochromatin Genome Project (5R01HG000747-14). Work by Rinaldo Bertossa was supported by the Division for Earth and Life Sciences (ALW) (grantnr. 817.02.020) with financial aid from the Netherlands Organization

for Scientific Research (NWO). Work by Márcia Bitondi was supported by Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP 05/03926-5), Brazil. Work by Seth Bordenstein was supported by grant NIH R01 GM085163-01. Work by Erich Bornberg-Bauer and Andreas Schueler was supported by the German science foundation, grant DFG BO2544/4-1. Work by Don Gilbert was supported by NSF DBI-064046 and NSF TeraGrid grants. Work done by the McClure group was funded by NIH/NIAID grant AI028309-13A2. Work by Bart A. Pannebakker and David M. Shuker was supported by the Natural Environment Research Council. Work in the Robertson lab was funded by USDA grant 2008-35302-18815. Work in the Romero-Severson lab was funded by the National Science Foundation (IOS-0432195) and the Indiana 21st Century Research and Technology Fund. Work by Saurabh Sinha was funded by NSF Grant DBI-0746303. Work by Arndt Telschow was funded by a Postdoctoral fellowship of the Volkswagen foundation. Work by Takuji Yamada, Jean Muller and Peer Bork was supported by BioSapiens (17191), BMBF grant Neuronet (17282) and EMBL. Sanger and Illumina genomic sequences are deposited in the NCBI trace archives (http://www.ncbi.nlm.nih.gov/Traces/home/) and NCBI Short Read Archives (http://www.ncbi.nlm.nih.gov/sites/entrez?db=sra), respectively, and can be accessed from the NCBI Taxonomy Browser database (http://www.ncbi.nlm.nih.gov/Taxonomy/) records for *N. vitripennis* (taxid= 7425), *N. giraulti* (taxid= 7426), and *N. longicornis* (taxid= 7427). The taxonomy database can be searched using either the species name or taxonomy ID. Gene sequences generated to examine intraspecific variation with the species are also deposited in Genbank (accession numbers: GQ896384-GQ896485, GQ904857-GQ905493). The alignment for the PRANC gene analysis is deposited in Treebase (www.treebase.org, ID SN4709).

Materials and Methods

I. Strains, Sequencing and Assembly

1. Strains for Genome Sequencing

All three sequenced Nasonia species routinely inbreed and have very low levels of intraspecific sequence variation (see intraspecific strain variation below). The use of highly inbred lines allows efficient assembly of these genomes because assembly is not complicated by extensive sequence polymorphisms often found in outbred diploid species. To even further reduce sequence heterogeneity in the strains used in genome sequencing, further inbreeding was conducted. The method for each strain is described below.

N. vitripennis: The AsymCX strain used in genome sequencing originated from the standard inbred laboratory strain LabII (*S1*). To produce an even more inbred line, a single virgin female was provided with hosts and then mated to her son. Six plus generations of sib mating then occurred, followed by 3 generations of sib matings during antibiotic curing (with rifampin) to produce a highly inbred strain free of the endosymbiont *Wolbachia* (*S2*). This strain was put into diapauses for approximately 1 year, removed from diapauses, then sibmated for three more generations and expanded for production of material for genome sequencing. DNA for genome sequencing was extracted from ~50 grams of mixed male and female pupae. There are no sex chromosomes in *Nasonia*, and therefore male and female DNA is the same.

N. giraulti: The strain RV2x(U) was derived from the isofemale inbred strain RV2 (*S1*). The strain for sequencing was created from an individual mother-son mating. Individual sibling matings were performed for 2 more generations before the strain was cured of *Wolbachia* using rifampin for 3 generations, which also involved sibmating. The strain was then maintained as a standard laboratory culture before being expanded for production of ~15 grams of pupae for DNA extraction. *N. giraulti* has extremely low levels of intraspecific variation, and therefore this further inbred strain was effectively isogenic.

N. longicornis: The strain IV7(U) was produced from the standard inbred strain IV7 (*S3*) by 3 generations of sib mating during rifampin treatment to eliminate *Wolbachia*. An isogenic line was then established from this strain. Approximately 30 grams of pupae were provided for DNA extraction to for genome sequencing.

2. Sequencing and Assembly

We sequenced the three *Nasonia* species *N. vitripennis*, *N. longicornis*, and *N. giraulti* using two different sequencing technologies. Inbred strains of each species that had been antibiotically cured of the endosymbiotic bacteria *Wolbachi*a (*S1*) were used for genome sequencing AsymCX for *N. vitripennis*, RV2X(U) for *N. giraulti*, and IV7(U) for *N. longicornis* (U indicates uninfected). *N.*

vitripennis was chosen as the reference species and sequenced to 6.2X genome coverage using Sanger technology on the Applied Biosystems 3730 platform. *N. giraulti* and *N. longicornis* were designated as comparative species, and these were sequenced to 1X sequence coverage on the Sanger platform, and 12X coverage in 45 bp fragment reads on the Illumina GA2 platform. For *N. vitripennis*, 3,108,554 successful sequence reads were generated and assembled into 295 Mb of ordered and oriented contigs using the Atlas assembly suite (*S4*) (see Table S1 for general statistics). This reference genome assembly (available from NCBI with accession number AAZX00000000.1) was used for gene model annotation and analysis.

We assessed the assembled *N. vitripennis* genome sequence for completeness and accuracy by comparison to 19 finished BAC sequences (Genbank accession numbers: AC185133, AC185134, AC185141, AC185142, AC185143, AC185288, AC185289, AC185290, AC185332, AC185333, AC185334, AC185335, AC185336, AC185337, AC185338, AC185339, AC185348, AC185349, AC185350, AC185351, AC185352) and 18,000 ESTs. 98% of the BAC sequences are represented within the assembled sequence. We estimated within contigs an error rate of 5.9×10^{-4}.

The 1X Sanger and 12X 45-bp-fragment Illumina data sets from both *N. giraulti* and *N. longicornis* were independently aligned to the *N. vitripennis* reference sequence using Mosaik assembler release 0.9.891 (2009-02-01; http://bioinformatics.bc.edu/marthlab/Mosaik). We used a hash size of 17 and a maximum mismatch percentage value of 0.1 (allowing a maximum of 10% mismatch over the length of the read). Custom perl scripts were used to output consensus genome and gene model sequences for the two sister species from the alignment ace files generated by Mosaik assembler. Consensus bases were determined by summing the qualities for all base calls at a single position in the total alignment: the base with the highest quality was used for the consensus. The consensus quality for each position in the sequence was determined subtracting the qualities of all other base calls from the quality of the consensus call. If the consensus quality after this procedure was less than 1, the consensus base call was changed to N. Areas where no sister species read could align to the *N. vitripennis* reference are represented by Ns in both the sister species consensus genome sequence and gene model files. 39.1% of *N. giraulti* and 45.0% of *N. longicornis* reads aligned to the *N. vitripennis* reference, which allowed 62.0% (109,425,004 Ns in 288,014,297 bp sequence) and 62.6% (108,422,994 Ns in 289,852,042 bp sequence) of the *N. giraulti* and *N. longicornis* genomes respectively to be assayed. General quality and coverage statistics are shown in Figures S10 and S11, respectively. Alignment is more reliable in more conserved parts of the genome, thus 84.7% (4,258,451 Ns in 27,878,327 bp coding sequence) and 86.3% (3,818,260 Ns in 27,873,753 bp coding sequence) of *N. vitripennis* coding sequence had alignment in *N. giraulti* and *N. longicornis* reads respectively. Additionally, we looked at coding exon sequences in the alignments and found that of 101,500 OGS v1.2 exons, 78,303 (77.1%) had full alignments in *N. giraulti* and 77,900 (76.7%) had full alignments in *N. longicornis*. At the gene level, we looked at 18,823 OGS v1.2 gene models,

of which 9,250 (49.1%) were complete with no Ns in the *N. giraulti* alignment. Similarly, 8,792 (46.7%) had complete alignments with no Ns in the *N. longicornis* alignment.

To estimate the quality of the aligned sequence, we compared the *N. giraulti* consensus aligned sequence to 3 finished BACs (Genbank accession numbers AC185140, AC185330, and AC185331) from an *N. giraulti* BAC library. Within the consensus aligned sequences we saw an error rate of 3.8×10^{-3}. As alignment quality is dependent on the conservation of the sequence being aligned to, we also looked at the quality of *N. giraulti* coding sequences from the OGS. We compared the *N. giraulti* OGS to the three finished *N. giraulti* BAC sequences. We identified 20 gene models with 163 exons totaling 41,709bp from the *N. giraulti* OGS represented in the three finished BACs. 40,788 bp were represented in the N. giraulti OGS sequence or 97.8% including 14 perfect gene models. Within the 40,788bp of aligned *N. giraulti* OGS there were 6 errors - an error rate of 1.47×10^{-4}.

II. Genome Annotation

1. Gene Models, Annotation, and Official Gene Sets

Only a handful of *Nasonia* genes had been sequenced prior to this project. Consequently, we heavily utilized automated gene predictions to aid our understanding of gene content in the wasp. The first Official Gene Set (OGS v1.0) was generated using NCBI's gene prediction pipeline (described below), and contained 2 subsets: RefSeq genes, which are supported by experimental data, and ab initio genes, which are not. The second OGS (v1.1), was generated using the RefSeq genes plus NCBI ab initio genes which showed homology in the comparative genome analysis (see global comparison of gene repertoire below). To generate the third OGS (v1.2), three additional gene predictors were run on the *Nasonia* genome assembly v1.0 assembly, including two that use a combination of transcript/protein alignments and ab initio modeling (AUGUSTUS, and Fgenesh++), and their outputs combined with the NCBI gene set (OGS v1.0) using GLEAN. Fgenesh++ and AUGUSTUS were two of the three best performing gene finders in the nematode genome annotation assessment project (nGASP) (*S5*); the NCBI pipeline was not evaluated in the nGASP project but performed similarly to AUGUSTUS in the current study. OGS v1.2 was constructed from a non-redundant combination of these predictions and additional predictions based on homology and EST support (described below) and is available for future genome-wide analyses of *Nasonia* gene content. Finally, we performed manual annotation and analysis for selected genes and gene families of particular biological interest. For this purpose a Chado database and Apollo instance was deployed to allow remote annotation by members of the *Nasonia* Genome Working Group.

a. NCBI's gene prediction pipeline

The NCBI gene prediction pipeline uses a combination of homology searching with ab initio modeling. cDNA, EST, and protein alignments to the genomic sequences are used as partial or complete support for particular splice patterns, representing an underlying preference to use experimental information whenever possible to support a particular gene model. cDNAs and ESTs were aligned to the genomic sequences using Splign (*S6*), and proteins were aligned using ProSplign (*S7*). Alignments were categorized into either coding sequence (CDS) or untranslated region (UTR) by identifying the best scoring CDS for the alignment (using the same scoring system used by NCBI's Gnomon ab initio prediction tool (*S8*), and marking any CDS scoring above a certain threshold as coding sequence. CDSs that lack a translation initiation or termination signal were categorized as incomplete. After determining the UTR/CDS nature of each alignment, the alignments were assembled using a modification of the Maximal Transcript Alignment algorithm (*S9*), taking into account not only exon-intron structure compatibility but also the compatibility of the reading frames. Two coding alignments were connected only if they both had open and compatible CDSs. UTRs were connected to coding alignments only if the necessary translation initiation or termination signals were present. There were no restrictions on the connection of UTRs other than exon-intron structure compatibility.

Transcript sequences used as evidence to support the NCBI annotation (NCBI build 1.1) included all *Nasonia* transcripts available as of July 2007, including approximately 60,000 ESTs from *N. vitripennis* and *N. giraulti*. Protein sequences used to calculate gene models included the complete genome annotations of human, *D. melanogaster*, *C. elegans*, and *S. cerevisiae* (derived from the RefSeq database), and all insect proteins annotated on cDNAs available in GenBank. An additional set of proteins including well-supported predictions from other species were used to refine some gene models.

The results can be divided into models with a complete CDS, including translation initiation and termination signals, and models with an incomplete or partially supported CDS. Incomplete or partially supported models were directed to Gnomon for extension by ab initio prediction. Alignments from the available *N. vitripennis* ESTs were used to optimize Gnomon ab initio prediction parameters. Additional proteins, including other insect genome annotations currently available in GenBank, are included in a second round of alignments to refine some gene models. Overlapping models with alignments supporting the entire CDS were combined into alternatively spliced transcript variants of a single gene. Models with a completely or partially supported CDS are considered to be of sufficient quality to be represented in the NCBI RefSeq database (*S10*). Models containing a debilitating mutation such as a frameshift or nonsense mutation were categorized as either transcribed or non-transcribed pseudogenes. A subset of pseudogenes are likely to be functional genes that have errors in the *Nasonia* genome assembly v1.0 and may be re-classified as protein-coding genes with subsequent improvements to the assembly and annotation. Gnomon was also used to predict pure ab initio models in regions of the genome that lacked any

cDNA, EST or protein alignments. A subset of Gnomon ab initio models are likely to be protein-coding genes but lacked supporting transcript or protein alignments to the genomic sequence at the time of the annotation run.

In total, the NCBI gene set (OGS v1.0) includes 9,159 protein-coding loci with complete or partial CDS support (the RefSeq set), 1,575 pseudogene loci which also have complete or partial CDS support but contain a frameshift or nonsense mutation within the CDS of the gene model, and 16,459 ab initio models that lacked supporting transcript or protein alignments (Table S19). Protein-coding models are represented in the RefSeq database with XM_ transcript and XP_ protein accession identifiers (*S10*), and include gene description records in the Entrez Gene database (*S11*). Most pseudogene models do not have RefSeq accessions, but are included in the Entrez Gene database. All models are available for search by BLAST at http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid=7425, and additional information can be found at http://www.ncbi.nlm.nih.gov/projects/genome/guide/wasp/.

b. AUGUSTUS

An initial training set was constructed by 1) clustering *Nasonia* ESTs, 2) aligning them to the genome and 3) by searching in the mRNA fragment produced by the alignment and the genome sequence for long (>= 300 bp) and likely complete (including stop codon and in-frame stop codon upstream of first ATG) ORFs. The parameters of AUGUSTUS (*S12*) were then trained and optimized for *N. vitripennis* on this training set of 831 genes. Besides sequence-intrinsic evidence, AUGUSTUS incorporated evidence ('hints') from EST alignments and protein alignments in the predictions. The ESTs were from *N. vitripennis* (18,299) and from *N. giraulti* (19,063). Protein homology information was derived by aligning proteins from UniRef50 and all Apocrita (wasps, ants, and bees) proteins from Genbank. AUGUSTUS was run without UTR predictions or alternative splicing.

c. Softberry FgenesH and FgenesH++ gene predictions

Fgenesh gene finder and Fgenesh++ pipeline (*S13*) have been applied to scaffold sequences to build *Nasonia* gene models. Both predictors used *Nasonia*-specific gene-finding parameters computed by running a special round of gene predictions in *Nasonia* sequences (with *Drosophila* parameters) and selecting for the training set those predicted genes that encode amino acid sequences having significant similarity with known proteins from various organisms (E-value less than 10-9 for the Blastp matches with the alignment covering at least 70% of the known protein length). Fgenesh++ is a pipeline for automatic prediction of genes that, in addition to Hidden Markov Model (HMM) based ab initio gene finder Fgenesh, includes software modules that can incorporate information from full-length cDNA, ESTs or known proteins to improve performance of ab initio gene finding. For *Nasonia*, we did not use cDNA or ESTs and just incorporated protein supported predictions using the eukaryotic part of the NCBI NR non-redundant protein database. There are 26,115

predictions (including incomplete gene structures) in 6,181 scaffold sequences. Among them 3,661 predicted genes with protein support and 22,454 ab initio predictions.

d. GLEAN consensus gene set

Individual gene prediction sets were integrated using GLEAN (*S14*). GLEAN is a tool for creating consensus gene lists by integrating gene evidence. It uses Latent Class Analysis to estimate accuracy and error rates for each source of gene evidence, and then uses these estimates to reconstruct the consensus prediction based on patterns of agreement/disagreement observed between each evidence source. GLEAN analysis labels each prediction with a confidence score reflecting the underlying support for that gene. GLEAN was run seven times using different combinations of the following gene prediction lists: NCBI RefSeq, NCBI Gnomon (excluding RefSeq), Fgenesh, Fgenesh++ (*S15, S16*), and Augustus (*S17*) as well as aligned Swissprot (*S18*) metazoan proteins and ESTs. The proteins were aligned to the genome assembly using Exonerate (*S19*) with a minimum 60% percent identity and 80% alignment coverage. ESTs from *N. vitripennis* and *N. giraulti* were aligned to the genome assembly using GMAP (*S20*) with 98% identity and 80% alignment coverage.

A "gold standard" set was created using 82 manually annotated coding sequences to evaluate the seven GLEAN sets and to compare them to input gene prediction sets. FASTA (*S21*) was used to align the gold standard sequences to the predicted gene models. To evaluate accuracy of intron/exon structure for each gene prediction set, the number of gold standard sequences with perfect matches to a predicted gene model was determined. A perfect match was defined as an alignment in which both sequences were completely aligned with at least 99% identity and no gaps. To evaluate completeness of a gene prediction set, the number gold standard sequences that matched a predicted gene model with at least 99% identity, not considering gaps or alignment coverage, was determined. In addition to alignments with the gold standard, each gene prediction set was evaluated for agreement with EST splice sites after aligning the ESTs to the genome assembly using Exonerate. The GLEAN sets were also compared to the RefSeq set, which was considered to be the most conservative set of gene predictions. It was noted that inclusion of some combinations of ab initio gene prediction sets in the GLEAN analysis resulted in high gene numbers, with a large fraction of gene models appearing to be split when compared to RefSeq gene models. Therefore the following parameters were considered in selecting the GLEAN set: (i) number of gene models (Table S20); (ii) number of high identity matches to gold standard sequences with complete alignment coverage (Table S20); (iii) number of high identity matches to gold standard sequences regardless of alignment coverage (Table S20); (iv) number of split gene models compared to the RefSeq set (Table S20), and (v) agreement with splice sites of aligned ESTs (Table S21). GLEAN6, the GLEAN set generated using RefSeq, Gnomon, Augustus, and Fgenesh was selected, because it did not perform worse than the best input dataset (RefSeq) in the gold standard comparison, it did not appear to have a high level of gene model splits

when compared to RefSeq, and it had a reasonable number of predicted genes compared to well annotated insects such as *Drosophila melanogaster*.

e. Official Gene Set (OGS) 1.2

In order to provide a comprehensive gene set for future analyses, we have also generated a non-redundant set of gene predictions designed to incorporate the most useful models from the various sets of predictions. The *Nasonia* official gene set (OGS v1.2) includes the following:

1) The subset of NCBI protein-coding gene predictions with complete or partial support (NCBI build 1.1) and incorporated into the NCBI RefSeq database (9,160 loci). These predictions have the benefit of being immediately available from NCBI with permanent accession numbers (designated with XM_ or XP_ prefixes) and are incorporated into the NCBI BLAST non-redundant databases.
2) The NCBI gene predictions classified as pseudogenes with complete or partial support (1,575 loci). These loci were included because a limited amount of manual curation indicated that a subset of the pseudogene predictions are likely to be functional protein-coding genes with defects (frameshifts and/or nonsense mutations) resulting from errors in the genome assembly. The nucleotide and protein FASTA sequences provided for the pseudogenes are modified to compensate for potential defects in the assembly sequence so that they represent the correct sequence if the defect is an error in the genome sequence. The pseudogene definition lines clearly indicate that they contain frameshifts and/or nonsense mutations, and it is advisable to verify the gene sequence and model structure to determine if the locus is truly functional in *N. vitripennis*.
3) Genes from the GLEAN set that do not overlap with the NCBI predictions in categories 1 and 2 (6,246 loci). These were included because the NCBI RefSeq set of supported genes is likely to be an underestimate of the true gene count in *Nasonia* due to the limited amount of transcript support available at the time of the analysis, and statistical support provided by the GLEAN consensus analysis suggested that these additional loci are likely to be functional genes.
4) A subset of NCBI ab initio predictions that do not overlap with categories 1, 2, or 3 but were found to have EST or protein homology support in later analyses (1,870 genes). Of the 27,431 Gnomon predictions, 8,505 were not retained in the RefSeq plus GLEAN6 combined gene sets. Among these, 1,870 gene models (http://insects.eugenes.org/arthropods/data/nasonia/notglean/) that lack transposon annotation are found to have significant protein homology or are supported by ESTs made available after the initial analyses. A homology criterion of $p < 10^{-5}$ with reciprocal matches was used. The EST analysis also yielded 185 novel full-length genes not contained in these predictions.

The OGS v1.2 can be downloaded from:
http://nasoniabase.org/nasonia_genome_consortium/datasets.html

2. Validation and Manual Annotation of Gene Sets

a. Comparison to additional EST sequences

An additional 115,140 ESTs sequenced from four *Nasonia vitripennis* cDNA libraries were made available after the initial annotations were computed, providing additional evidence to support gene predictions and a dataset suitable to validate the gene sets. PASA (*S9*) was used to assemble cDNA-gene models from all available *N. vitripennis* and *N. giraulti* ESTs as of November 2008 (Table S22). A PASA database was constructed (http://insects.eugenes.org/arthropods/data/nasonia/pasa_est/) for *Nasonia* that provides web access to EST assembly summaries and details, EST validation and correction reports for gene predictions, providing a useful reference for expert gene annotators.

PASA includes methods to check gene predictions against EST assemblies, and annotate discrepancies, such as where one EST assembly joins two gene predictions (a "gene merge"). Not only does it tabulate these errors, but it provides expert annotators with a web report on each gene model discrepancy, and will produce output of corrected gene models. The six gene prediction sets for *Nasonia* were run through the PASA EST validation along with the official final set (OSG v1.2), with results in Table S23. Validation statuses are listed (see also Fig. S12), along with gene prediction totals, sensitivity and specificity calculated from EST results. Statuses includes "Incorporated" (no change), "UTR added" where EST extends non-coding regions, common since most of these predictors did not call UTR regions, and "Alternate splice" where ESTs support multiple splice forms for a gene. Model errors are noted with "Gene extension" where the coding region is extended by ESTs, "Internal gene structure rearrangement" where predicted coding exons do not match EST exons, and "Gene merge" where two genes are joined by one EST assembly.

The seven gene prediction sets performed similarly, with some variation in types of correct or incorrect models. Specificity relates the portion of total gene models supported by ESTs, ranging from 33% for the larger prediction sets to 65% for the smaller predictions. This needs to be balanced by Sensitivity, which shows the portion of ESTs found by a predictor, ranging from 68% for the small set, to 77% for the largest. Overall, 54.7% of OGS v1.2 gene models were found to have support from a PASA EST cluster. Of those models, 62.8% were consistent with the transcript data, whereas 37.1% had discrepancies suggesting the gene model may need to be refined. These statistics are based on non-curated EST data which may not accurately reflect every gene structure, but the overall gene-level specificity and sensitivity scores are similar to the better gene prediction sets in the nGASP evaluation of gene prediction software (*S5*). Planned updates to the RefSeq and OGS gene sets will incorporate changes supported by the additional EST evidence.

b. Gene model correction for missed or fused genes

Potentially missed genes (not in the NCBI XP_ or ab_initio gene set) were identified from orthologous groups which showed an ortholog present in most

other species (*P. humanus*, *A. mellifera*, *T. castaneum*, *D. melanogaster*) except *N. vitripennis*, and where the *A. mellifera*/*T. castaneum* ortholog gave a significant TBLASTN hit to the *N. vitripennis* genome (Table S24, available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html). We used Genewise/Fgenesh+ predictors followed by manual refinement to suggest the best possible gene model for 68 of such *Nasonia* genomic regions. Notably, for six cases it proved impossible to build a respectable gene model due to gaps, stop codons, or other problems; for ten cases the predictions are likely truncated or interrupted because of the gene start/end or an internal exon is in the gap of the current genome assembly; twelve predictions overlap exons of other gene models that probably need to be corrected; and 19 predictions are in introns of other genes. Furthermore, we used the orthology evidence to refine 95 gene models that were initially fused by the automated prediction pipeline. These updates will be incorporated into future versions of the OGS, and are available separately for download (Table S25, available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html).

c. Manual gene annotation

A database and web-based interface previously used for the *A. mellifera* and *T. castaneum* genome projects was configured at BCM-HGSC to allow members of the *Nasonia* Genome Working Group to provide additional annotation information for gene models, including nomenclature, orthology, and supporting evidence for characterized loci. This interface was also used to submit sequence information for some incorrect or missing gene models. Annotation data was submitted for 1,043 genes, including 71 novel genes and 260 with updated sequence information.

Some genes required more complex manual annotation that was performed using the Apollo Annotation editor (*S22*). Apollo can use either a RefSeq accession or a scaffold id to connect directly to the NasoniaBase (http://nasoniabase.org) and retrieve multiple types of evidence, including gene predictions from NCBI RefSeq, NCBI ab initio, Fgenesh++ and GLEAN, alignments of protein homologs from Swissprot metazoa, honey bee (both official gene set and ab initio), *Drosophila melanogaster* (from FlyBase) and *Tribolium castaneum* (from RefSeq), alignments of EST/cDNA from *N. vitripennis*, *N. girault*i, yellowjacket, braconid and fire ant, predicted pseudogenes, repetitive elements and transposable element homologs. These evidence sets are viewed as tracks in Apollo and used to correct existing gene models or create new gene models. Specific focus was placed on the verification of splice sites, identify start and stop codons, add UTRs or annotate additional splice variants, depending on available evidence. Completed annotations were saved as Chado-XML files, then uploaded to NasoniaBase. Thirteen annotators from the research community used Apollo to annotate 73 protein-coding gene loci, with a total of 84 unique transcripts.

Data from both manual annotation approaches is available for download at NasoniaBase

(http://nasoniabase.org/nasonia_genome_consortium/datasets.html) and will be incorporated into future versions of the OGS.

III. Gene Expression

1. EST Library Generation

Two non-normalized EST libraries were generated for each of *N. vitripennis* and *N. giraulti*, one each for late larval stages and one each for pupal and adult stages. Total RNA extractions were performed using RNeasy Mini Kit (Qiagen Inc., Valencia, CA) and cDNA libraries were constructed and sequenced as previously described (*S23*). This effort produced 27,553 and 30,770 total reads for *N. vitripennis* and *N. giraulti*, respectively. However, we noticed a high level of chimeric sequences present in the library, where fragments of multiple genes were erroneously joined together into a single sequence. To remove these chimeric ESTs, we ran BLAST searches (*S24*) of the 100 bp on both ends of each sequence against the *N. vitripennis* genome. ESTs with ends that matched different scaffolds (contiguous genomic sequences) of over 40 Kb in length were removed. Additionally, we removed all ESTs that matched mitochondrial sequences, resulting in a total of 38,848 high quality ESTs. Following an assembly, 12,990 cDNA clones that most distinctly represented unsequenced extreme 3' of 5' ends of gene transcripts were arrayed to produce an additional 5,924 ESTs for a total of 44,772 cDNA sequences. (Accession numbers: ES613911-ES651267).

2. Construction of a Normalized cDNA Sequencing Resource

cDNA library construction: Male and female pupae and whole bodied adult males and females were harvested to create four full-length enriched normalized cDNA libraries for sequencing. Total RNA was isolated using Trizol reagent (Invitrogen) and was subsequently purified using the RNeasy protocol (Qiagen). The cDNA libraries were produced using the Creator SMART (Clontech) system by following the manufacture's instructions. After the cDNA synthesis but prior to cloning, the cDNA pool was normalized using the Trimmer-Direct cDNA normalization kit (Evrogen), amplified then ligated into the pDNR-LIB vector. The vector-cDNA ligants were bacterial transformed into TOP10 competent cells (Invitrogen), grown onto selective 2xYT agar plates overnight and individual colonies were archived by freezing within 15% glycerol 2xYT selective media. These libraries are available to the research community by the Indiana University Center for Genomics and Bioinformatics. Sequencing reactions were performed by priming at the 5' end of cDNA using vector primer pDNRlib30-50 (TAT ACG AAG TTA TCA GTC GAC G) and by priming at the 3' end using vector primer M13rev (AAA CAG CTA TGA CCA TGT TCA C) with ABI BigDye chemistry and the 3730xL sequencer. Vector and poor quality sequences were trimmed from the sequencing reads and ESTs were assembled into contigs using ESTPiper

(*S25*). EST sequences have been deposited in Genbank, Accession numbers: GE352825 - GE467204.

Assembly of the ESTs: Including 44,772 ESTs from the previous sequencing effort (described above), the ESTPiper program assembled 131,966 ESTs out of 159,746 sequences that passed quality assurance thresholds. The assembly to the *Nasonia* genome sequence scaffolds began first by using BLAT (*S26*) to find overlapping and mate-paired EST clusters, then by using PASA (*S9*) to merge sets of compatible overlapping EST alignments to identify alternative splice variants. The following parameter options were applied: blat min. identity = 90%; blat max. intron = 750 Kb; clustering min. coverage = 80%; clustering min. overlap = 40 bp; clustering max. magnification = 10 bp.

3. Tiling Expression Microarrays

NimbleGen tiling arrays: We used NimbleGen high-density 2 (HD2) arrays for transcriptome investigations. The custom 4-array (chip) set consisted of 8.4 million isothermal long-oligonucleotide probes that are 50-60 nt in length and that span the *Nasonia* genome sequence at overlapping intervals of 33 bp, on average. Each slide contained 27,000 markov model random probes that are not represented in the genome for setting background level thresholds. All probes were designed using NimbleGen's ArrayScribe software and the quality assurance tests of the probes were conducted by IU-CGB in-house algorithms. Signal to background ratios were determined by first calling probes that fluoresced at intensities greater than 99% of the random probes' signal intensities; therefore only 1% of fluorescing probes should be false positives. The arrays reliably produced high signal to background ratios; $\log_2$ ratios of eight were observed for signal over background.

We conducted three replicates each using RNA from independent biological extractions of male and female early embryo (0-10 hrs), late embryo (18-30 hrs), 1st instar larvae, and pupae. Additional experiments were performed comparing transcription in testis and the female reproductive tract. Samples were prepared at 25º C as follows: Approximately 200 *N. vitrirpennis* (AsymCX) virgins were collected as black pupa. When eclosed, half of them were provided with males and were allowed to mate overnight, and the other half were kept as virgins. Virgin setting result in all-male progeny whereas mated females produced ~89.5% female progeny under the design used here. Females were provided 15-20 *Sarchophaga bullata* hosts in groups of 20 females for 24 hrs to induce production of eggs. The hosts were then removed and females were left overnight (~18 hrs). To collect embryos, individual females were given access to a host at one end (to restrict the oviposition site) and allowed to lay eggs for 6-10 hrs before being removed. Embryos were then harvested immediately (early embryos), 18 hrs later (late embryos), or 51 hrs later (1st instar larvae). All embryos and larvae were collected in an RNAse free environment. The host was cracked open and the "cap" removed to expose embryos. Dissecting needles were used to gently scrape embryos from the surface of the host and transfer them into a 1.5 ml tube pre-chilled on dry ice. Samples were stored at -80º C. If at anytime the host was punctured or embryos were exposed to host

hemolymph, they were discarded. Estimates of the number of embryos per replicate (three per life stage/sex) were recorded; early embryos ranged from 300-900, late embryo 140-500, 1st instar larvae 245-520. Since sex cannot be determined at larval stage, some of the mated female hostings were allowed to mature to adulthood then males and females were counted to determine the sex ratio. Early larvae showed an average of 82.9% females and late larvae had an average of 84.2%. Pupae collections were made among the progeny of mated females provided with hosts for 48 hrs. They were sorted by sex and stage (early yellow, red-eye, half black, and black pupae). Equal numbers (*S2*0) of pupae from each stage were then pooled prior to RNA extraction. Males and females were extracted separately. Male reproductive tracts (testes, seminal vesicles, and accessory glands) were dissected in phosphate buffered saline (PBS) from and transferred with a small amount of PBS into a tube on dry ice. A total of 60 testes per replicate were dissected (10 each yellow/red, salt & pepper, and black pupal stages). Female reproductive tracts (30 per replicate) were removed from 1-3 days post eclosion virgin females and transferred to a tube on dry ice.

Tissue was disrupted and homogenized using Trizol reagent (Invitrogen) and extracted RNA was purified using the Qiagen RNeasy protocol with optimal, on-column DNase treatment from specific tissues. Beginning with at least 0.5 µg of total RNA (for early to late embryo) or at least 1.0 µg (for other tissue types), a single round of amplification using MessageAmp™ II aRNA kit (Ambion) produced between 30 and 45 µg of cRNA for embryo RNA and greater than 100 µg for all other tissue types. Starting with 10 µg of cRNA, double strand cDNA synthesis was carried out using the Invitrogen SuperScript Double-Stranded cDNA Synthesis kit using random hexamer primer followed by DNA labeling using 1 O.D. CY-labeled random nonomer primer and 100U Klenow fragment (3>5 exo) per 1 µg double-stranded cDNA. The use of random primers ensured that all transcripts hybridize to the array, which contains probes designed solely from a single strand of the DNA sequence. Both sexes for each tissue type were alternatively labeled and a dye-swap was included among the replicate experiments. Duel-color hybridization, post-hybridization washing and scanning were done according to the manufacturer's instructions. Images were acquired using a GenePix 4200A scanner with GenePix 6.0 software. The data from these arrays were extracted using the software NimbleScan 2.4 (Roche NimbleGen).

Array Data Analysis: The data analysis was performed using the statistical software package R (http://www.r-project.org/) and Bioconductor (http://www.bioconductor.org/). The signal distributions across chips, samples and replicates were adjusted to be equal according to the mean fluorescence of the random probes on each array. All probes including random probes were quantile normalized across replicates (*S27*). Scores were assigned for each predicted exon of OGS v1.2 genes, for each sample, based on the median $\log_2$ fluorescence over background intensity of probes falling within the exon boundaries. The exons were deemed to be transcribed only when greater than ½ or their tiled length was expressed. Genes validated by tiling array or EST data are shown in Table S26, available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.

Analysis of transcript-form diversity: The tiling array experiments allowed detection of up to ten transcript forms from single gene models (maximum of one form per sample-type hybridized on the arrays). We counted a total of 37,009 transcript variants from 14,924 genes represented on the arrays. To create a validated set of transcript forms, 114,885 EST sequences were aligned via BLAST ($p < 10^{-100}$) against the predicted cDNA sequence for each detected transcript. Transcripts were unambiguously supported if one or more ESTs aligned uniquely to a variant. Similarly, if one or more ESTs aligned to more than one transcript-form of a gene, it was reported as ambiguously supported.

We compared the number of transcript-forms for *Nasonia* and *Drosophila melanogaster* by first obtaining a set of previously computed orthologs identified by OrthoDB (*S28*). 7,719 *Nasonia* genes were confidently assigned to 9,800 *D. melanogaster* orthologs within 6,650 ortholog groups. For our analysis, these were classified into sets of 1) one-to-one orthologs that were single copy genes in both species; 2) one-to-many orthologs that were single copy genes in *Nasonia*, yet duplicates were located in *Drosophila*; 3) many-to-one othologs that specified *Nasonia* paralogs related to single copy *Drosophila* genes; and 4) many-to-many orthologs that related multi-copy genes in both species. Transcription-form data for *D. melanogaster* were obtained from FlyBase (Release dmel_r5.16 mRNA) (http://flybase.org/).

4. MicroRNA Prediction

Four different sets of computational microRNA predictions were generated by two different groups. Both groups created homology-based sets (Sets 1a and 2a) using slightly different strategies. The two groups also produced miRNA predictions based on sequence conservation: Set 2a was based on comparison with 40 animal species, and set 2b was based on comparison between *N. vitripennis* and *A. mellifera*. An additional set (set 3) was generated using bioinformatics analysis of 454 sequencing reads from small RNA libraries of *N. vitripennis*.

a. miRNA prediction set 1a: First homology-based prediction set

All known metazoan miRNAs from miRBase 12.0 (*S29*) were aligned to the *Nasonia vitripennis* genome assembly v1.0 using BLASTN of the WU-BLAST package [BLASTN 2.0MP-WashU (http://blast.wustl.edu)] with the following parameters adapted for cross species comparison: -M 1 -N -1 -Q 3 -R 2 -W 9 -filter dust -mformat 2 -hspsepSmax 40 -e 1e-3. BLASTN matches longer than 20 bp were extended at both ends to match the length of the query sequence. In a following step the extended blast hits were aligned to their query sequence using MAFFT (*S30*) with the following parameters: --maxiterate 1000 --localpair -quiet. To remove unstable or spurious hits a set of features were calculated for each hit and evaluated: 1) total sequence length > 40 bp, 2) 100% conserved seed (nt 2-8 of the putative mature part) region in regard to the query sequence, 3) more than 90% sequence identity for the mature part, 4) sequence conservation of the total precursor sequence larger than 60%, 5) no more than two gaps in the mature

part, 6) minimum free folding energy smaller -15 kcal/mol, 7) more than 40% of the bases should be paired, 8) mature regions should not overlap a multi-branch loop, 9) RandFold p-value smaller 0.05 if precursor conservation smaller 95% to any known miRNA. RandFold (*S31*) estimates the stability of the folding compared to dinucleotide shuffled folded sequences (100 randomizations). As the query set was redundant (e.g. containing dme-bantam, ame-bantam) the final predictions were clustered according to their locus on the *Nasonia* genome using GALAXY (*S32*). From a single locus the hit with the highest conservation of the mature miRNA and the highest overall percent alignment identity over the entire putative pre-miRNA was used as a single representative sequence. Using this approach, we predicted 51 microRNAs in *Nasonia* (Table S5 available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html).

b. miRNA prediction set 1b: Comparative sequence approach based on SVM

We used a comparative approach on the basis of a Support Vector Machine (SVM) model of hairpin-like structures followed by an orthology assignment step. This method allows prediction of novel miRNAs that do not show sequence homology to known miRNAs. The complete method is described in (*S33*) and the results are available from http://cegg.unige.ch/nasonia_genome. What follows is a brief outline of the basic principles: First, an ab initio SVM model was created to score stem-loop like sequences extracted from the genomic sequence with RNAfold (*S34*). Second, an orthology assignment pipeline grouped putative precursors from over 40 animal species, then precursors within groups were aligned. In a third step the orthologous groups were again subjected to an SVM model designed to distinguish alignments of orthologous miRNA sequences from other ncRNA alignments or false positive predictions, taking into account typical conservation patterns in pre-miRNA sequence alignments. Those predictions were put forward as the miRNA prediction set 1b if they had an ortholog in at least the bee genome and were supported by a 454 sequence read (described under miRNA prediction set 3 below). In total this method yielded three new *Nasonia* miRNAs which are conserved in the bee genome and supported by a high-throughput read, but not present in set 1a.

Using a 10 Kb window 17 out of the 54 *Nasonia* miRNA from combined miRNA prediction sets 1 and 2 group in cluster containing two or more miRNAs. Four clusters contain two miRNAs, one cluster three and the biggest cluster contains 6 miRNAs within a 1.9 Kb long region (nvit-mir-71; nvit-mir-2a, nvit-mir-13a; nvit-mir-13b; nvit-mir-2b; nvit-mir-2c). All *Nasonia* miRNA in this cluster are on the forward strand, whereas in the honey bee the miRNAs of this cluster are all on the reverse strand. Although the synteny is conserved, in honey bee the miRNAs span a larger chromosomal region (2.7 Kb) than in *Nasonia*. Except the six miRNA in this region, an alignment of the honey bee region vs. the *Nasonia* region does not show any significant sequence conservation (Fig. S13).

c. miRNA prediction set 2a: Second homology-based prediction set

Mature miRNAs and their respective stem-loop precursors were downloaded from miRBase (*S29*). Each precursor miRNA sequence was aligned against Nvit1.0 with WU-BLAST (http://blast.wustl.edu). BLAST searches were performed by seeding only the mature region of the precursor miRNA to minimize false positives, and then allowing seed extensions outside the mature region. Sequences corresponding to each BLAST hit were extracted from the assembly, extending the extracted sequence to the length of the original query. A global alignment between query (precursor miRNA) and subject sequence (extracted region) was constructed with T-COFFEE (*S35*), and the number of substitutions was determined. The free energy of folding of the subject sequence was computed with RNAfold (*S34*). A PRSS analysis between the two sequences was performed with 1,000 iterations in order to assess the statistical significance of the alignment and confirm that the two sequences were homologous. PRSS, which is part of the FASTA sequence comparison package (*S36*), is used to construct local alignments between a query and a database of shuffled subject sequences to generate a distribution of alignment scores, which is used to compute an E value for the alignment of the query to the actual subject. In our case, the query was the precursor miRNA and the subject was the extracted region of the assembly. Putative miRNA homologs were kept if they were at least 95% identical to the known miRNA or if the following conditions were met: 1) similarity score of at least 65% throughout the entire global alignment, 2) free energy of folding $\leq$ -20 kcal/mol, and 3) PRSS score $\leq 10^{-05}$. This pipeline identified 45 microRNAs.

d. miRNA prediction set 2b: Predictions based on conservation between wasp and honey bee

In order to identify possible new microRNAs shared between the two hymenopteran genome assemblies (*N. vitripennis* Nvit_1.0 and *A. mellifera* Amel_4.0), we looked for regions of perfect conservation between the two genomes. Using ultraconserved regions as starting point, we developed a strategy to score candidate regions for their potential to encode new microRNAs.

Genomes were aligned using WU-BLAST with the following parameters: B = 100000, V = 100000, spoutmax=0, nogaps, Hspsepmax = 100, filter = seg, mformat=2, hspmax=0, W=20, N=9000. Aligned *Nasonia* sequences were extracted and extended by 65 nt at each end. Redundancy was minimized by removing every sequence that was a perfect substring of another sequence of equal or longer length. Pairwise alignments of the honey bee-wasp pairs were constructed using the Smith-Waterman algorithm as implemented in the program water of the EMBOSS package. Pairwise alignments were then evaluated using RNAalifold (*S34*) in order to determine the secondary structure and the free energy of folding of each pair of sequences. Only sequence pairs having the canonical stem-loop precursor found in microRNAs were further considered. Sequence pairs were then evaluated using RandFold (*S31*) with 1,000 iterations per sequence. Sequence pairs were scored on the basis of their alignment score, deltaG of folding and RNAfold score. Alignment scores and deltaG of folding

were normalized to sequence length. Thirteen microRNAs that were not detected in set 2a were predicted after scoring each sequence pair using the following formula:

> Pair Score = (Alignment length / alignment score) + (1/((RandFold score in *Apis* + RandFold score in *Nasonia*)/2))/200) + (ABS (deltaG of folding for the pair)/alignment length)*10)

The scoring function was developed this way in order to give equal weight to the different parameters evaluated on the candidate sequences. These 13 predictions were supported by comparison with sequences from the small RNA library (described under miRNA prediction set 3 below).

e. Compilation of Computational miRNA Analyses

The results of the four computational analyses (sets 1a, 1b, 2a, 2b) were combined. Predicted miRNAs were considered to be a single locus if genome coordinates overlapped and they were on the same DNA strand. The union of all sets resulted in 61 unique miRNA loci. The union of sets 1a and 2a consisted of 53 unique loci with known homologs. Four of the set 2b predictions (based on wasp/bee ultraconserved sequences) overlapped with homology-based predictions from set 1a, and two overlapped with SVM based predictions from set 1b. Therefore, a total of 9 novel miRNA were identified in the computational analyses (union of predictions without known homologs from sets 1b and 2b). Eleven of the 61 miRNAs were considered to be hymenoptera specific, either because they were identified only using wasp/bee ultraconservation (set 2b) or a known homolog only existed in bee.

f. miRNA prediction set 3: Small RNA library sequencing and analysis

Sample Types and Preparation: Samples for low molecular weight (LMW) RNA library production were prepared by the Werren laboratory using the genome sequenced strain AsymCX and sent to the laboratory of J. Evans for RNA extraction and library construction. Dissections were performed over ice using a sharp scalpel and each specimen was immediately placed on dry ice and then into a -80° C freezer prior to shipment on dry ice. Dissection instruments were changed between sample types. The following sample types were provided: <u>whole pupae</u> (YP♂ = 39 male yellow pupae; YP♀ = 39 female yellow pupae; SPP♂ = 38 male salt & pepper pupae, SPP♀ = 36 female salt & pepper pupae; BP♂ = 56 male black pupae; BP♀ = 37 female black pupae), <u>female dissected parts</u> (H♀= 61+50+53 adult female heads dissected at three different times, T♀= 63+63+60 adult female thoraces dissected at three different times; A♀= 60+50+57 adult female abdomens dissected at three different times), <u>male dissected parts</u> (H♂ = 57+54+14 adult male heads dissected a three different times; T♂ = 56+58+15 adult male thoraces dissected at three different times; A♂ = 47+54+15 adult male abdomens dissected at three different times).

Low molecular-weight RNA extraction: RNA extractions were carried out on the following tissue pools: 1) 39 male yellow pupae, 2) 39 female yellow pupae, 3) 38 male salt & pepper pupae, 4) 36 female salt & pepper pupae, 5) 56 male

black pupae, 6) 37 female black pupae, 7) 53 female heads, 8) 50 female thoraces, 9) 57 female abdomens, 10) 54 male heads, 11) 58 male thoraces, 12) 54 male abdomens. MicroRNA was extracted from each tissue pool following the mirvana microRNA extraction protocol (Ambion) with a final volume of 100µl. 70µl of this was ethanol precipitated and resuspended in 11µl.

Purifying 17-27mers from LMW RNA: Each of the twelve samples were prepared for loading by adding an equal volume of 2X formamide buffer containing 0.05% xylene cyanol FF (XC) and 0.05% bromophenol blue (BPB) to the RNA. Samples were mixed and denatured at 65º C for 5 min. then spun and maintained on ice. The entire solutions containing RNA were loaded on a prepare 15% TBE-Urea Polyacrylamide Gel (Invitrogen), along with 10-bp DNA ladder (Invitrogen) and 5' RNA control primer for size. The gel was run at 200 V until good size separation was achieved. The bands were visualized by staining with Sybr Gold dye (Molecular Dynamics), 10µl in 100µl TBE buffer. Gel bands representing the 20-27-nt fragment were isolated and transferred to a 1.5ml tube. The gel slice was crushed with a pestle and two volumes of RNA elution buffer (0.3 M NaCl) was added. The gel was incubated overnight at room temperature with shaking. The following day, the RNA was extracted in one volume chloroform and ethanol precipitated in a final volume of 6.5 µl ddH_2O.

5' Adaptor Ligation and Purification: For 5' adaptor ligation, 11 µl reaction mixes were prepared that consisted of: 6.5 µl purified 20-27 nt RNA, 0.3 µl of 200 µM PAGE-purified 5' adaptor (5' GGU CUU AGU CGC AUC CUG UAG AUG CAU 3'; de-protected and desalted by gel filtration), 1X Ligation Buffer, 10U T4 RNA Ligase (Ambion), 40U RNase OUT (Invitrogen). The reaction was incubated at room temperature for 4 hrs and terminated by the addition of 2X formamide buffer (as previous). The sample was heat denatured and run on a 10% TBE-Urea Polyacrylamide Gel (Invitrogen) and stained with Sybr Gold dye. The corresponding 49-56 nt bands were isolated in two volumes RNA elution buffer and extracted overnight (as previous) with a final volume of 6.7 µl ddH_2O.

3' Adaptor Ligation and Purification: For 3' adaptor ligation, 11 µl reaction mixes were prepared that consisted of: 6.5 µl purified 5' ligation product, 0.3 µl of 200 µM PAGE-purified 3'Adaptor (5' pAUG CAC ACU GAU GCU GAC ACC UGC idT 3'; p = phosphate; idT = inverted deoxythymidine), 1X Ligation Buffer, 10U T4 RNA Ligase (Ambion), 40U RNase OUT (Invitrogen). The reactions were incubated at room temperature for 4 hrs and stopped by adding 11 µl 2X formamide Loading Buffer. The sample was heat denatured and run on a 10% TBE-Urea Polyacrylamide Gel and stained with Sybr Gold dye. The corresponding 73-80 nt bands were isolated in two volumes RNA elution buffer and extracted in one volume of chloroform followed by ethanol precipitation. The RNA pellet was suspended in a final volume of 20 µl ddH_2O.

RT-PCR of adaptor-ligated small RNAs: To eliminate any contaminating DNA, the RNA was treated in a 10 µl reaction consisting of 8 µl adaptor-ligated RNA (0.2 µg/µl), 10U DNase (Ambion), 1X DNase buffer, and 20U RNase OUT (Invitrogen). The reaction was incubated at 37º C for 1 h, then 75º C for 15 min. cDNA was synthesized using 10 µl DNase-treated adaptor-ligated RNA (0.16 µg/µl), 10 pmol RT- primer (5' GCA GGT GTC AGC ATC AGT GT 3'), and 2 mM

dNTP. The reactions heated to 42º C for 2 min. followed by the addition of 0.5 µl Superscript II Reverse Transcriptase (Invitrogen), 1.5 µl Superscript buffer, 2 µl DTT, for a final reaction volume of 15 µl. The reaction was incubated at 42º C for an additional 50 min. followed by 70º C for 15min.

PCR amplification of cDNA: Each sample was first amplified using specific linker tags to identify tissue type, as PCR products will be pooled for 454 runs. There were four forward linker primers as follows: 454LinkerA (5' GCC TCC CTC GCG CCA TCA G**AA CAC** GGT CTT AGT CGC ATC CTG TA 3'); 454LinkerB (5' GCC TCC CTC GCG CCA TCA G**AC GTA** GGT CTT AGT CGC ATC CTG TA 3') ; 454LinkerC (5' GCC TCC CTC GCG CCA TCA G**AT CTG** GGT CTT AGT CGC ATC CTG TA 3'); 454LinkerD (5' GCC TCC CTC GCG CCA TCA G**CA GCA** GGT CTT AGT CGC ATC CTG TA 3'). All reactions used the same reverse primer: 454RTLinker (5' GCC TTG CCA GCC CGC TCA GGC AGG TGT CAG CAT CAG TGT 3'). For amplification, 50 µl reaction mixes were prepared that consisted of: 2 µl cDNA, 0.4 µM each primer, 1.5 U Taq DNA polymerase (Roche), 1X buffer, 0.1 µl master amp (Epicentre Biotechnologies), and 0.4 µl 10 mM dNTP. The following cycling parameters were used on the PTC-100 thermal cycler (MJ research): 95º C for 5min. then 15 cycles of 94º C for 40 sec., 68º C for 4min. PCR products were run on a 15% TBE-Urea Gel and bands around 100 bp were isolated and extracted (as previous). The cleaned DNA was resuspended in 25 µl ddH_2O. 2 µl of this cleaned PCR product was used as the template for further PCR amplification using the same chemistry and cycling conditions. In addition, the products from this second PCR were run on a 15% TBE-Urea Gel with 18-30 nt bands being isolated, eluted and cleaned with chloroform and ethanol precipitation, as previous. The final product was resuspended in 6.5 µl ddH_2O.

Cleaned PCR Product with 454 linkers (*A-D)*: Male Black Pupa 061107-1 20 µl (A): n=56 male black pupae; Female Black Pupa 061107-2 20 µl (B): n=37 female black pupae; Female Head 061107-3 20 µl (A): n=50+53 pooled female heads; Female Thorax 061107-4 20 µl (B): n=50+50 pooled female thoraces; Female Abdomen 061107-5 20 µl (C): n=50+57 pooled female abdomens; Male Head 061107-6 20 µl (A): n=57+54+14 pooled male heads; Male thorax 061107-7 20 µl (B): n=56+58+15 pooled male thoraces; Male Abdomen 061107-8 20 µl (C): n=47+54+15 pooled male abdomen; Male Yellow Pupa 082307-1 20 µl (A): n=39 male yellow pupa; Female Yellow Pupa 082307-2 20 µl (B): n=39 female yellow pupae; Male Salt & Pepper Pupa 082307-3 20 µl (C): n=38 male salt & pepper pupae; Female Salt and Pepper Pupa 082307-4 20 µl (D): n=36 female salt & pepper pupae.

Bioinformatics Analysis of Small RNA sequences: Sequences generated by 454-sequencing were trimmed of linkers, and sequences ≥ 16 nt after trimming were used in further analysis (44,344 sequences, Table S27, available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html). RNA sequences were mapped to the *Nasonia* genome assembly v1.0 using Megablast (*S37*) with a word size of ten. Alignments of 19 to 23 nt were identified, and those with redundant or overlapping coordinates were clustered together to provide 2,490 representative alignments. The aligned sequences and

their reverse complements were extracted from the assembly along with 75 nt in each direction resulting in a larger window of ~172 nt. RNAfold (*S34*) was used with default settings to determine minimum free energy (MFE) scores for overlapping 110 nt windows. The window with the lowest MFE became the candidate precursor microRNA. BLASTN was used to identify and remove sequences that were homologous to other cellular small RNAs such as tRNA, rRNA, URNA, and snoRNA from Rfam 9.1. Candidates that overlapped coding sequences of the *N. vitripennis* official gene set v1.2 by at least 70% were removed, resulting in 1,397 unique candidates. 900 candidates overlapping repetitive elements (see methods for repeat analysis) were removed, reducing the set to 497 candidates. RNAfold was used again with a sliding window, this time with a varying frame size (50-100 nt). Candidates were retained if 1) folds contained single-branched stem loops, 2) MFE was less than -15, and 3) the mature sequence (expressed small RNA) was located in the stem. This resulted in 86 predicted microRNAs, 27 of which were also identified in computationally predicted miRNA sets. The 900 candidates overlapping repetitive elements were also processed through RNAfold with similar parameters, resulting in 672 candidates that overlapped repeats. While these 672 were not included in the final set 3 predicted miRNA count, their existence suggests potential involvement in control of repetitive elements.

g. Combining small RNA library and computational miRNA prediction sets

The prediction sets were combined by identifying loci with overlapping coordinates in the genome assembly. A total of 120 unique miRNA loci were identified in the combined computational and small library datasets. Thirty-nine of the computational miRNA predictions were supported by alignment to small RNA sequences, yet only 27 of the miRNAs identified in bioinformatics analysis of the small RNA library (set 3) overlapped with computational predictions (Table S6). This difference is likely due to differences in parameters used in sequence trimming and alignment, suggesting that criteria used in bioinformatics analysis of the small RNA library were stringent, and that 120 is a conservative estimate for the Nasonia miRNAome.

IV. Mapping of Scaffolds and Genotyping

1. Crossing Scheme

To map and orient scaffolds of the *Nasonia* genome assembly v1.0, we generated a mapping population of 112 haploid F_2 hybrids. By analyzing haploid F_2 hybrid males, we took advantage of the haplodiploid genetics that the *Nasonia* model system offers to efficiently collect marker segregation data. The mapping population was initiated by crossing virgin females of *N. vitripennis* with males of *N. giraulti*. The emerging diploid and genetically identical F_1 hybrid females were then raised as virgins to make them lay unfertilized eggs. Due to the haplodiploid sex determination in *Nasonia*, all eggs developed into (haploid) recombinant (F_2

hybrid) males. We further took advantage of the available genome sequences for *N. vitripennis* and *N. giraulti* by using the same strains for the cross experiment that had been sequenced (i.e., AsymCX and RV2X(U)). Both stains are cured of their *Wolbachia* endosymbionts and are highly inbred. The genetic uniformity of the strains allowed us to directly apply single-nucleotide polymorphism (SNP) identified between the available genome sequences for high-throughput genotyping the haploid recombinant F_2 hybrids in the mapping population with the Illumina and Nimblegen genotyping arrays (see below). To minimize pseudolinkage effects (i.e., the apparent linkage between markers/genomic regions due to epistatic interactions causing mortality), 16 hrs old F_2 hybrid male embryos were utilized because this stage is prior to any F2 hybrid mortality (*S1*).

2. Illumina SNP Genotyping Array

For high-throughput genotyping of *N. vitripennis* × *N. giraulti* hybrids, we developed a Illumina SNP genotyping array with 1,534 markers. Single-nucleotide polymorphism was identified by aligning trace sequences of *N. giraulti* with the ortholog scaffold sequences of *N. vitripennis*. To reduce the number of false positives, we considered only potential SNPs for which at least two *N. girauti* trace sequences were found that indicated a SNP difference between *N. vitripennis* and *N. giraulti* and that otherwise both had identical sequences 100 bp up and downstream of the potential SNP. Given the limited capacity of the array, only a subset of the potential SNP markers was considered and selected to maximize the number of scaffolds represented on the array. Details of the Illumina SNP genotyping array design are published by (*S38*). All SNP genotyping experiments were performed following the Illumina Goldengate Genotyping Assay protocol provided by the manufacturer (Illumina). Each sample was hybridized to the Illumina Sentrix Array Matrix platform, scanned with an Illumina Beadstation 500 reader, and analyzed with the Illumina Beadstudio software.

3. Nimblegen Mapping Array

In addition to the Illumina SNP genotyping array, we developed a custom Nimblegen microarray that utilizes clusters of SNPs to create strong hybridization differences between the corresponding *N. vitripennis* and *N. giraulti* alleles. All scaffolds were screened for 80 bp sequences which contained 5-15 SNPs and/or indels between *N. vitripennis* and *N. giraulti*, at least 2X coverage in *N. giraulti*, no intraspecific polymorphisms, and all SNPs having a PHRED (*S39*) quality score >= 50. Oligopicker 2.3.2 (*S40*) was then used to select 70 bp oligonucleotides from the resulting sequences, and all sequences which matched the genome sequence in multiple places or the *N. vitripennis* mitochondrion were removed. These oligonucleotides were then placed on a custom Nimblegen microarray. The array contains paired oligonucleotides for ~19,000 loci, which encompass 929 scaffolds and 86% of the sequenced genome. This array can be used to genotype individuals or quantify the amount of *N. vitripennis* vs *N. giraulti* DNA for each locus in bulk DNA samples.

4. Scaffold Mapping

We considered genotype data from the Illumina genotyping array plus data of additional 111 SNP, length polymorphic, and present/absent markers that had been collected for various reasons in different labs to infer a high-density recombination map of the *Nasonia* genome. Details of the *Nasonia* high-density recombination map and its inference are summarized by (*S38*). This map covers 265 scaffolds (64% of the assembled genome).

5. Mapping Genes and Genome Features of Interest

We used the *Nasonia* high-density linkage map (*S38*) to determine the position of 36 genes of interest on the map. Specifically, genes involved in DNA methylation (DNA cytosine-5-methyltransferases *Dnmt1a*, *Dnmt1b*, *Dnmt1c*, *Dnmt2*, *Dnmt3*); nuclear genes interacting with mitochondrial gene products and or involved in ATP generation (*Atp5O*, *AtpD*, *Cox5A*, *Nd23, AmpK, NdUbr*); genes involved in sex determination (*dsx* = doublesex, *tra* = transformer, *fru* = fruitless); developmental genes (*admp*, *bambi*, *dpp* = decapentaplegic, *gbbA* = glass-bottom-boat, *gbbB* = myostatin, *mav* = maverick, *MedA* = medea, *MedB*, *otd1* = orthodenticle-1, *putA* = punt, *putB*, *spzA* = spaetzle, *spzB*, *tkv* = thickveins, *TollA*, *TollC*, *TollD*, *tsgA* = twisted gastrulation/crossveinless-like, *tsgC*, *vnd* = ventral nervous system defective); immunity genes (*Pgrp-LC*); and eye-color genes (cinnabar = *cinn*). We further used the information of mapped scaffolds to study and illustrate gene content across the *Nasonia* genome. We inferred gene density by estimating the gene content for windows of approximately 3.2 cM on the linkage map using the approach described by (*S38*) and analyzing the sequence data of the 265 mapped scaffolds, the latter representing 63.6% of the total sequenced genome. For this purpose, we analyzed a subset of the *Nasonia* OGS v.1.2 that had been filtered for transposable elements (see further below).

6. Mapping Visible Markers

Positions of visible markers in the sequenced genome were estimated by mapping, using a combination of between-species introgression lines and F2 recombination screening. Introgression lines were created of the wildtype *N. giraulti* (RV2Xu or R16A strain) alleles of *N. vitripennis* mutants by 8+ generations of backcrossing into the *N. vitripennis* (mutant) background. Introgressions of the *N. giraulti* alleles of eye color mutants *rdh-5*, *bk-424* and *st*-DR were created and made homozygous. The locations of a *N. giraulti* introgressed quantitative trait locus affecting male wing size, *wdw* (*S41*), was also identified. Genomic location of the introgressed *N. giraulti* sequence was determined using the competitive genome hybridization microarray described below.

Finer resolution mapping of the eye color "*R-locus*" on Chromosome 5 (including allelic *st-DR* and *pe-333* eye color mutants) was achieved by backcrossing the wildtype *N. giraulti* alleles of *pe-333,pu* into the *pe-333,pu* mutant background. F2-F5 males recombinant between *pe-333* and *pu* were

collected and genotyped with PCR using length-polymorphism markers designed to differentiate *N. giraulti* and *N. vitripennis* (Table S28). Mapping of *st-318* and *mm* on Chromosome 3 was achieved similarly.

V. Genome Feature Analyses

1. Repetitive DNA

a. De novo repeat library construction

We generated de novo transposable element (TE) libraries for the *Nasonia vitripennis* genome using PILER-DF (*S42*). PILER-DF searches for repeat regions that are alignable at least three times in the genome and thus provides a relatively unbiased way to estimate the number of repeat families irrespective of taxa. Thus, it is by their virtue of having at least three globally alignable copies in the genome that these elements are classified as TEs. Briefly, the release 1.0 version of the *Nasonia* genome was self-aligned using PALS. Regions represented greater than three times in the genome were extracted by PILER-DF as putative repeats. All PILER-DF predictions were self-aligned using MUSCLE (*S43*). From each family of repeats we used custom scripts to identify a single centroid representative that was defined as the most closely related to all other sequences. We further simplified our repeat library by removing all predictions that were more than 90% similar to another prediction over 90% of their length. This dataset was then curated using RepeatMasker (*S44*), Tandem Repeats Finder (*S45*), and BLAST comparison to transposable element domains (i.e. gag, pol env, integrase downloaded from PFAM (*S46*). BLAST alignments scoring greater than 50 bits were used to curate putative transposable elements. We also used custom Perl scripts to predict tandem inverted duplications and long terminal repeats at the terminal ends of PILER-DF predictions. Predictions were more than 20% tandem repeat content, as judged from TRF4 results, were annotated as novel tandem repeats while predictions with similarity to any known repeats or protein domains were curated as being similar to that repeat (i.e. LTR, LINE, Gypsy, etc.). PILER-DF predictions were screened against known genes from *D. melanogaster* and *A. mellifera* to screen out false positive predictions. A tab-delimited annotation of the PILER-DF repeats and a FASTA library of the Repbase and annotated PILER-DF predictions used for this study can be found in Tables S29 and S30 available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.

b. Gene and genome masking

We used our de novo PILER-DF predictions along with species-specific Repbase repeats to mask the *N. vitripennis* genome using RepeatMasker 3.17-3.25 (Tables S31 and S32 available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html). We also used BLASTN with default parameters to map nine known repeat elements that

have been shown to be localized to B chromosomes (Psr105-3, 18-1, 22-2, nv79-16a, 85-7, 126-6, 104-6, and AAAGTCT(T/C)GACTT) (Table S33) (*S47*).

We sub-divided the large scaffolds in the existing release 1.0 assembly of the *Nasonia* genome into 100,000 bp segments so that we could better observe the distribution of repeats across the genome. We segmented the 6,181 scaffolds in the *Nasonia* assembly into 8,459 segments with lengths of 100,000 bp or less. We used Repbase (*S48*) and our PILER-DF predictions as libraries to mask the genome using RepeatRunner (*S49*) and RepeatMasker (*S44*) as above. A frequency histogram of scaffold repeat content was generated (Fig. S1A) and an arbitrary cut-off for putative heterochromatic regions was designated the point where the frequency of low repeat content scaffolds is greater than 1% and high repeat content scaffolds is less than 1% frequency (Fig. S1B). Empirically this region is the part of the distribution that is to the right of where the majority of mapped sequence is found (0-15% repeat content, Fig. S1) and was determined empirically by measuring the repeat content of cytologically-verified heterochromatin in *D. melanogaster*. We used custom Perl scripts to tabulate the nucleotide composition of unmasked and masked genome segments to use for frequency histograms. We also used our repeat library to screen the RefSeq supported gene annotations, GNOMON gene predictions, and GLEANR consensus gene models for potentially repetitive ORFS that may not represent *bona fida* coding genes. All genes in the *Nasonia* OGS v1.1 were masked using RepeatMasker 3.25 with either custom PILER-DF repeat libraries or the default RepBase *N. vitripennis* library as well as TRF4 (Table S34 available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html).

c. Annotation of retroid elements

The Genome Parsing Suite (GPS) (*S50*) is a radical departure from the well-known RepeatMasker, which is used to mask out and count all repetitive elements using consensus DNA sequences (*S44*). All methods that use DNA as the query suffer from the loss of signal due to mutational saturation. The GPS is the first method to use protein sequences and amino acid motif identification, rather than DNA, in the search for Retroid agents, thus providing a deeper query into genomes. While the GPS software can be populated with any set of homologous protein sequences, in the results presented here it was populated with 90 reverse transcriptase (RT) sequences to assess all potentially active and ancient autonomous Retroid agents in the wasp genome. Sixty-four of these queries are representative of all Retroids identified in all Eukaryotes to date and allows clear classification between Retroid types (i.e., retroviruses, retrotransposons, retroposons etc.). In addition, 26 *N. vitripennis* specific queries were identified in the initial GPS analysis and added as queries in the final run for more refined classification.

Briefly, the RT protein sequences are used as WU-tBLASTn (*S51*) queries to provide the data parsed by the GPS. Stage I GPS sorts and filters raw WU-tBLASTn hits, which are redundant and contain false positives, due to: 1) alternative alignments for a given query to a specific region, 2) cross coverage of the queries, and 3) counting as unique, a number of small hits that are actually

from the same gene. After sorting by query, chromosome, polarity and reading frame the GPS compounds small hits, and removes false positives due to cross coverage on these compounded hits. The GPS removes redundancy by deleting hits that are completely covered by a longer hit to the same position, thereby preventing overestimation of the amount of potential RT genes.

Stage II of the GPS extends each RT gene's position in the *N. vitripennis* genome by 7 Kb both upstream and downstream, creating a 14 Kb cutout. Using this RT-outward approach, the GPS is able to construct the new Retroid agent genomes. This 14 Kb sequence is sufficient to encompass most newly identified Retroid agents, plus any insertions the sequences may have. WU-tBLASTn is used a second time to compare each 14 Kb cutout with a query component. For example, if the unique RT hit was pulled out by the N_vitripennis_Pao-like retrotransposon (NV_Pao), then only the NV_Pao component library is used to identify the rest of the genes. If a RT hit is ambiguous between multiple queries, then each of those queries' component libraries are searched, and the highest scoring over all components is labeled the closest neighbor to the new Retroid agent. If the sequence has all the gene components in the query's genomic order without any inserts it is considered full length. Unlike some methods, the GPS does not define a full-length genome by long terminal repeats (LTRs) or untranslated regions (UTRs) alone. It is known that many LTR or UTR bounded Retroid genomes have insertions and/or deletions (indels) within these boundaries and, therefore, are not full length as is the query. The GPS classifies LTR/UTR bounded genomes as true full-lengths and various subsets of indel mutants relative to each query Retroid genome. Part of the GPS workflow is to assess for new LTRs not present in the query component libraries. We have compared available LTR-detection methods and LTR_FINDER (*S52*) performs the best with our data to detect the presence of new LTRs.

2. Heterochromatin Gene Annotation

For orthology analyses, 150 cDNA-supported *D. melanogaster* heterochromatin genes were compared to the *N. vitripennis* RefSeq and ab initio gene sets (not including GLEANR predictions) using TBLASTN (*S24*). We identified orthologs to known *D. melanogaster* heterochromatin genes (*S53*) using reciprocal BLAST analysis (*S24*). *N. vitripennis* orthologs to *D. melanogaster* heterochromatin genes were hand annotated using the Apollo curation tool (*S22*). Briefly, provisional gene models predicted by the Baylor Genome Center pipeline were compared to aligned cDNA sequences, gene predictions, and BLAST alignment data. Intron-exon boundaries were refined to match supporting data and the repeat content of the surrounding 100,000 bp or smaller regions was assessed (see above). Similarly, 100,000 bp or smaller segments from *N. vitripennis* with repeat content greater than or equal to 16% were identified and several hundred genes were annotated. These represent the putative heterochromatin genes in *N. vitripennis* and have been deposited in nasoniabase.org.

2. Telomere Annotation

Several attempts were made to identify the telomeres of *Nasonia* from the genome assembly and raw reads, but all failed. The identity of the *Nasonia* telomeres therefore remains enigmatic. A single copy gene encoding a seemingly intact and reasonably full-length telomerase gene is partially present in the assembled genome and was repaired using raw reads from all three species. The gene spans three contigs. An upstream non-coding exon was only identified because of overlapping 5' and 3' ESTs from cDNA clone 96O19 from female pupae (GenBank GE393116.1 and GE394043.1). This exon is at least 214 bp long with no obvious ORF. It is followed by a 3.5kb intron in *N. vitripennis*, which spans from the 5' RC end of 1.4 kb scaffold 2765 (Ctg16865 in GenBank AAZX01016116.1) to around 357 kb on the reverse strand in 1.6 Mb scaffold 25 (Ctg13580 in GenBank AAZX01013059.1). The 106 bp gap between these two contigs is covered by a single *N. vitripennis* read. Presumably then, scaffold 2765 belongs within the inter-contig gap at this location within scaffold 25. This long intron contains a ~3kb sequence repeated commonly in the *N. vitrpennis* genome that is absent from the *N. giraulti* intron. The first coding exon begins near the 3' end of 2.8 kb Ctg13580. The first 58 bp of this exon upstream of the candidate translation start AUG codon contains at least one stop codon in each reading frame, so assuming that this cDNA is an accurate representation of the 5' end of the cDNA, there is no scope for additional N-terminal coding capacity. This is important because the resultant protein is missing approximately 140 amino acids at the N-terminus when compared with the *Apis mellifera* telomerase (*S54*), and most vertebrate telomerases. While this might appear devastating, the silkmoth *Bombyx mori* and red flour beetle *Tribolium castaneum* telomerase proteins also appear to have similarly short N-termini (*S54, S55*). The first 144 coding bp of this exon are present in the contig, but the remaining 390 bp of this exon, a 100 bp intron, the 308 bp second coding exon, and the first half of a 291 bp intron were built from a combination of two non-overlapping *N. vitripennis* reads that extend the flanking contigs, and single reads from *N. giraulti* and *N. longicornis* that bridge a ~100 bp gap between the ends of these two reads. The rest of the gene continues with six exons separated by short introns in the first 2 kb of 18 kb Ctg4043 (GenBank AAZX01003975.1), which continues on the reverse strand of scaffold 25 to around 351kb. The resultant 697 amino acid predicted protein, except for the shortened N-terminus noted above, aligns well with the other insect and other animal telomerases. This composite version of this 7 kb gene, with introns in lower case, and the encoded protein, are provided in Figure S14.

In light of the presence of a seemingly intact and reasonably full-length telomerase or TERT gene, whose encoded protein should be capable of synthesizing telomeric repeats on the ends of telomeres, we searched the genome assembly for the TTAGG telomeric repeats identified in diverse insects and other arthropods (e.g. (*S56, S57*), and which are present in the telomeres of the silkmoth *Bombyx mori* (*S58)* and the honey bee *Apis mellifera* (*S54*). BLASTN searches revealed only two to four copies of TTAGG repeats internally in a few contigs, which are unlikely to be parts of telomeres. The red flour beetle

is unusual in having TCAGG telomeric repeats (*S55, S59*), so we also searched for that, and all other single base substitutions from TTAGG, without finding any significant stretches of repeats that could be telomeres. In case the telomeric repeats had failed to assemble entirely (which seems unlikely given the documented presence of a large variety of repeats in this genome - see main text), these searches were also performed on the raw reads without additional success. Searches for TTAGGG and related repeats, which are the more ancestral form in other animals (*S57*), to which *Nasonia* might have "reverted", were also unsuccessful. It appears, therefore, that while *Nasonia* has a telomerase gene, it does not have typical insect TTAGG or related telomeric repeats at its telomeres, although this conclusion might be invalid if it has telomeric repeats that have diverged by more than a single nucleotide from TTAGG.

The telomeres of the silkmoth *Bombyx mori* (see (*S58*) for review) and the red flour beetle *Tribolium castaneum* (*S59*) have many copies of a family of retrotransposon inserted into their TTAGG repeats. These SART family transposons are non-LTR retrotransposon as their "poly-A" tails usually end within a TTAGG repeat after the GG. Searches of the *Nasonia* genome assembly and the raw reads with a AAAAAAAAAAAAAAAAAAAAAAAAGGTTAGG query did not yield any worthwhile matches, so this kind of retrotransposon arrangement does not appear to be present in *Nasonia*. Similar searches in the *Bombyx* and *Tribolium* genomes yield abundant matches and extended examination of these reveals they are mostly to these kinds of retrotransposons in telomeres.

4. DNA Methylation and CpG Patterns

a. DNA methyltransferases

Annotation of Nasonia Dnmt orthologs: Honeybee and human DNMTs were used as query sequences to search (with BLASTP) the *Nasonia* RefSeq and ab initio databases for DNA methyltransferases. Pairwise alignments and percent identity and similarities (positives) between two DNMT sequences were determined using BLASTP pairwise alignment and Clustal W was used for multiple sequence alignment and for generation of the phylogram shown in Figure S15. The methyltransferase domain of each DNMT was identified using Pfam HMM criteria for DNA methylases. N-terminal domains wer identified using Pfam and NCBI Conserved Domains.

Confirmation of gene models: Gene models were tentatively confirmed using RT-PCR (Fig. S16) with primer pairs that flank predicted exon-intron boundaries. Total RNA was prepared separately from male and female yellow pupae of either the BI or the AsymCx strain using TRIzol Reagent (Invitrogen); cDNA was generated using Superscript 1st strand system (Invitrogen). A two-step cycling protocol was used with the annealing temperature close to or slightly above the primer Tms: 94º C 2 min; 35 cycles: denature at 94º C for 30 sec anneal/extend at 68º C 3 min followed by 68º C 3 min. RNA prepared from male and from female pupae gave the same results in these experiments. If follow-up analysis

was required, RT-PCR products were cloned using Invitrogen's PCR-4 TOPO vector system. Plasmid preparations were sent to Nevada Genomics for dideoxy sequence analysis. The following primers were used in the RT-PCR analysis of Dnmt1a: Dnmt1a LE4 CGA CTC AAT AAT CCT AAC GCT GCC GTT T LE4; Dnmt1a RE5 GTT GGA AAT GAG ATA GGG GTT CAC TGC C RE5; Dnmt1a RE6 CCG CTT ACA GTT GGA ATG TGA GCC ATA C RE6; Dnmt1a RE7 ATG AGG TAG ACA CCA GGG GAT CAG TGT G RE7; Dnmt1a RE8 ATT CGC GCA CAC TTA CGA CTC TTG TTT G RE8.

Phylogenetic analysis of DNMTs: Sequences were obtained for *Homo sapiens*: (*S60*); *Drosophila melanogaster*: (*S61*); *Anopheles gambiae*: (*S62*)*; Bombyx mori*: (*S63, 64*); *Apis mellifera*: (*S40*); *Tribolium castaneum*, (*S64*); *Daphnia pulex:* (*S65*); *Acyrthosiphon pisum*, *Pediculus humanus* and *Ixodes scapularis:* Arthropod Species Genomes Database http://insects.eugenes.org/arthropods/blast/ and http://insects.eugenes.org/arthropods/data/summaries/arthropod-gene-GCandCpG.html and Table S58 (Available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html).

Source information on divergence times: The approximate time of the protostome/deuterostome split is from (*S66*). Divergence times of *Nasonia* from other taxa in the protostome phylogeny were estimated as described below; all others are from Figure 1 of the *Apis* genome paper (*S67*).

Dating of the split between Nasonia and Apis: Despite continuing uncertainty about relationships among major hymenopteran lineages (*S68*), a reasonable time estimate for the divergence between *Nasonia* and *Apis* would be at least 160 MY, probably closer to 190 MY. We therefore use an estimate of 180MY in this figure. Fossils assignable to Aculeata (which contains *Apis*) and Chalcidoidea (which contains *Nasonia*) are known back to 130-140 MY (*S69*), so each lineage must be at least that old. Other distinct parasitoid lineages also date back this far based on both fossil and molecular divergence time evidence (*S70, S71*). Moreover, likely relatives of the Chalcidoidea, the Proctotrupoidea, are known from fossils about 160-180 MY old, and some questionable aculeate fossils are known back to about 150 MY (*S69*). Hymenopteran fossils in the broad sense are known back to more than 200 (maybe as old as 230) MY.

Dating of the split between Nasonia and other Holometabola: Given the recent finding from genomic comparisons that Hymenoptera are the earliest-diverging among the extant holometabolous insect orders (*S72, S73*) the divergence times between *Nasonia* and each of *Tribolium*, *Bombyx*, *Anopheles* and *Drosophila* should be the same, estimated at roughly 290-300 MY. A popular alternative view, that Hymenoptera are more closely related to the panorpoid complex of orders rather than more basal in divergence than Coleoptera, is also supported by some evidence and if accepted, would imply that *Nasonia* and *Bombyx*, *Anopheles* and *Drosophila* share a common ancestor more recently (roughly 260 MY) than does *Tribolium* (*S69*). But current evidence, including a recent study based on analysis of a large number of nuclear genes (*S74*), more strongly supports the Hymenoptera-basal position (*S68*). In addition, this recent study suggests that the Hymenoptera/other Holometabola split may have occurred

even earlier, as much as 350 MYA (*S74*), although that timing is not supported by fossil evidence at this point.

Dating of the split between Nasonia and Exopterygota: The split between Holometabola and the Paraneopteran orders, including *Rhodnius*, *Acyrthosiphon* and *Pediculus*, has been estimated at roughly 300 MY (*S69*), or just slightly older than the oldest divergences within Holometabola. Permian fossil deposits display an astonishing array of insect orders, only some of which survived the end-Permian extinction just under 250 MY. Current evidence suggests that much of winged insect ordinal diversity arose between 330 and 300 MY, a relatively short period of time compared to the long time since (*S69*). It is thus not surprising that relationships among many insect orders have been difficult to resolve even with significant effort from both comparative morphologists and molecular systematists (*S75*).

Dating of the split between Nasonia and Daphnia: After decades of uncertainty and dispute concerning higher arthropod relationships, current evidence seems to be definitively resolving the sister-group of Hexapoda to be among the Crustacea (*S76*). The origin of Pancrustacea has been estimated at about 600 MY based on molecular evidence (*S76*), while clear fossil evidence of Arthropoda does not appear until 60 MY later. In any case, the common ancestor between *Daphnia* and *Nasonia* would be more recent than this (probably close to but under 500 MY), since *Daphnia* belongs to one of the crustacean clades close to Hexapoda (*S77*).

Dating of the split between Nasonia and Ixodes: It is currently uncertain how old the split between Hexapoda and Chelicerata (including *Ixodes*) is, but it is clearly early Cambrian to Precambrian in age, or in the vicinity of 600 MY (*S76, S77*). Relationships of Chelicerata to other possibly basal arthropod lineages are still being explored (*S76, S78*)

b. Ratio of observed to expected CpG in GC content domains

We partitioned the genomic sequences into segments by the binary recursive segmentation procedure, DJS, proposed by (*S79*). In this procedure, the chromosomes are recursively segmented by maximizing the difference in GC content between adjacent subsequences. The process of segmentation was terminated when the difference in GC content between two neighboring segments is no longer statistically significant (*S80*). The distribution of RefSeq gene models within GC content domains was determined. Observed to expected CpG ratio was computed for individual GC content domains, introns and coding exons of RefSeq gene models. For each gene model, intron and coding exon O/E CpG were computed after concatenating introns or coding exons, respectively. Results are shown in Figure S2.

c. Normalized CpG content of *N. vitripennis* genes

The 'normalized CpG content' ($CpG_{O/E}$) for each gene is defined as $CpG_{O/E} = P_{CpG}/(P_C*P_G)$, where P_{CpG}, P_c and P_G are the frequencies of CpG dinucleotides, C nucleotides, and G nucleotides, respectively (estimated from each gene). We

calculated CpG O/E for 8,825 RefSeq genes, considering both genes (exons plus introns) and introns only.

d. Experimental test of DNA methylation in selected *N. vitripennis* genes

To choose candidate genes for analyses of DNA methylation, we first analyzed the level of CpG dinucleotide depletion of *N. vitripennis* genes. For each of *N. vitripennis* gene with RefSeq support (8,825 genes total), we calculated the metric 'normalized CpG content (CpG O/E)', defined as $P_{CpG}/(P_C x P_G)$ where P_{CpG}, P_C, and P_G correspond to the frequencies of CpG dinucleotides, C and G nucleotides of each gene (including exons and introns), respectively. It has been shown that in animal genomes, regions with high CpG O/E tend to be hypo-methylated while those with low CpG O/E are likely to be hyper-methylated (*S81, S82*). We chose three genes with relatively low CpG O/E values and two genes with relatively high CpG O/E values as candidates of direct test of DNA methylation. Gene names, CpG O/E values, and the primers used for analysis of DNA methylation are shown in the Table S35.

To quantify the level of DNA methylation of selected *N. vitripennis* genes using direct sequencing following bisulfite conversion. Total genomic DNA was isolated using Puregene DNA isolation kit (Gentra/Qiagen) from male and female *N. vitripennis* separately. DNAs (~500 ng) were bisulfite converted with the EpiTech Bisulfite conversion kit (Qiagen) following the manufacturer's instructions.

Bisulfite sequencing primers were developed using Methyl Primer Express Software (v1.0) (Applied Biosystems). Each primer pair was amplified in a 25 μl reaction using a PCR program of initial denaturation at 95° C for 10 minutes, then followed by 40 cycles of denaturation at 94° C for 30 seconds, annealing at 54° C for 30 seconds, extension at 72° C for 1 min., and a final step of extension at 72° C for 10 min. Amplified bisulfite PCR products were purified using the QIAquick Gel Extraction Kit (Qiagen), then cloned into pCR 2.1 vector by use of a TOPO-TA cloning system (Invitrogen) and transformed into TOPO10 chemically competent *E. coli* (Invitrogen). A least five positive clones were randomly selected for automated sequencing from male and female *N. vitripennis,* respectively.

We found evidence of DNA methylation from all five genes. However, the frequencies of methylated cytosines varied among genes (Fig. S3; Table S35).

5. Protein Domain Annotation: Procedure and Comparative Domain Analysis within a Five Proteome Comparison Set

We annotated 17 arthropod proteomes consisting of *Drosophila melanogaster* (r5.11)*, D. simulans* (r1.3)*, D. sechellia* (r1.3)*, D. yakuba* (r1.3)*, D. erecta* (r1.3)*, D. ananassae* (r1.3)*, D. pseduoobscura* (r1.3)*, D. persimillis* (r1.3)*, D. willistoni* (r1.3)*, D. mojavensis* (r1.3)*, D. virilis* (r1.2)*, D. grimshawi* (r1.3), *Apis mellifera* (OGS r2)*, Anopheles gambiae* (P3.49)*, Aedes aegypti* (L1.49)*, Tribolium castaneum* (51906) and *Daphnia pulex* (1) using HMMER2 (*S83*) and domain models obtained from Pfam A (*S46*). All sequences not covered by this annotation step were subject to annotation using rpsBlast and ADDA profiles

(*S84*). We used two different E-value cutoffs levels. Level 1 requires 10^{-5} and 10^{-9} for rpsBlast, while level 2 requires 10^{-3} and 10^{-7} for matches to Pfam A and ADDA profiles respectively. While all genome wide comparisons were done using the level 1 cutoff, we used the level 2 cutoff for quality checks. For the sake of simplicity the global comparative analysis was carried out only for a comparison set of four selected genomes (*D. melanogaster, A. mellifera, A. aegypti*, and *A. gambiae*) which represent the tree well with respect to *N. vitripennis* and are also well annotated. The global properties are summarized in Table S36. Level 1 domain annotation of the comparison set is available under http://adb.uni-muenster.de.

VI. Comparative Genomics

1. Global Comparison of Gene Repertoire

We defined orthologous relations among genes in *Nasonia* (using the 27,403 NCBI predicted genes in OGS v1.0, including both RefSeq and ab initio models), *A. mellifera* (honeybee), *T. castaneum* (beetle), *D. melanogaster* (fruitfly), *P. humanus* (body louse), as well as *D. pulex* (Daphnia) and *H. sapiens* (human). The orthologous groups were automatically inferred from all-against-all protein sequence comparisons (*S28*) using the Smith-Waterman algorithm, followed by clustering of best reciprocal hits from highest scoring ones to 10^{-6} E-value cutoff for triangulating BRH or 10^{-10} cutoff for unsupported BRH, and requiring a sequence alignment overlap of at least 30 amino acids across all members of a group. Furthermore, the orthologous groups were expanded by genes that are more similar to each other within a proteome than to any gene in any of the other species, and by very similar copies that share over 97% sequence identity. The second OGS (v1.1) was constructed from the RefSeq genes plus the ab initio models which showed homology in this analysis.

2. Synteny between *Nasonia* and Honeybee

To identify orthologous gene arrangements between *Nasonia* and honeybee (and from them to body louse for a sanity check, which was chosen for being most compact and having lower evolutionary rate than Diptera) we considered only 1:1:1 orthologs (~ 6,000) and required at least two such orthologs to be nearby on the same scaffold and not allowing more than one other intervening ortholog (out of these 6,000) to define synteny blocks. There are 1,346 such blocks spanning 4,042 1:1 orthologs, with the biggest block with 22 genes. A rough quantification of the level of such micro-synteny is the percentage of orthologs grouped into synteny blocks relative to the total number of considered orthologs.

3. Kegg Analysis

To identify potential functionally coordinated losses, we mapped all putative gene losses to KEGG (pathways and modules) and Biocarta pathways via human or fly orthologs and estimated the statistical significance of the coordinated losses using hypergeometric test (Tables S37 and S38). P-values for significantly enriched ($p < 0.01$) pathways are in bold. Gene losses in selected pathways were checked manually by using TBLASTN to the *Nasonia* genome and an E-value cut-off of 10^{-40}.

4. Gene Family Expansions/Losses

Groups of orthologous genes were delineated as described earlier (*S28*). Runaway expansions in *Nasonia* were defined as orthologous groups with multiple copies in *Nasonia* but otherwise single copy. Orthologous groups present in Human and *Nasonia*, but not *Drosophila*, were filtered requiring the BLASTP E-value of the human protein to *Nasonia* to be at least 10^5 lower than to *Drosophila*. Orthologous groups uniquely shared between *Nasonia* and human (not in any other examined insect) were required to have a TBLASTN E-value at least 1,000-fold smaller between human and *Nasonia* than to any other insect genome. Annotation was retrieved from ensembl via the fly or human ortholog. Hymenoptera-specific orthologs were filtered by requiring the Bee protein to be at least 1,000 closer (E-value) to the *Nasonia* ortholog compared with any *Drosophila* and *Tribolium* protein. In addition, TBLASTN comparisons were performed using the bee proteins as a query against the transcripts of a third hymenopteran species, the fire ant *Solenopsis invicta* (clusters and putative transcripts from http://fourmidable.unil.ch) (*S85*). Results are shown in Table S41 and Tables S39 and S40 available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.

5. Lateral Gene Transfers

The PRANC-domain, which occurs thirteen times in the *N. vitripennis* proteome, has so far only been described in pox-viral proteomes of *Chordopoxvirinae* (vertebrate pox-viruses). We used HMMER2 and the Pfam definition of the PRANC domain (PF09372) to scan the entire GenBank protein non-redundant database (release: 02.04.2009) to search for other eucaryotic proteomes with PRANC-domains. This scan yielded 262 significant hits (E-value $< 10^{-3}$), 237 of which correspond to pox-viral proteins. The other hits correspond to proteins from three different alpha proteobacteria of the order Rickettsialles, *Orientia tsutsugamushi* (strain Ikeda (13 hits) and strain Boryong (5 hits)), *Rickettsiella grylli* (1 hit), and two *Wolbachia* (*w*Ana (1 hit) and *w*Ri (1 hit)), from the parasitoid wasp *Cotesia congregata* (1 hit) and 5 hits in *N. vitripennis* (the other 8 had not yet been included in GenBank).

The Pfam Hidden Markov Model (HMM) describing the PRANC domain has been trained on pox-viral sequences only. To alleviate the bias towards pox-viral PRANC-sequences and increase the model sensitivity for other possible eukaryotic PRANC instances, we aligned all 262 PRANC-sequences identified in

the GenBank scan and used this alignment to train a new HMM with HMMER2. As pox-viruses are overrepresented in the alignment, we determined the 50 most informative pox-viral PRANC sequences from the alignment using T-Coffee (*S35*) and used these together with all non pox-viral sequences to train a non-biased HMM. Using this HMM, we were able to identify 318 significant hits (E-value < 10^{-3}), which include all 262 hits from the previous scan. All additional hits correspond to pox-viral proteins.

Primers were designed based on the sequence of the PRANC domain gene found in *w*Ana and *w*Ri (WRi_006810). (WolPRANC70F 5′-GATGCTAAGAGGGTTACAAGCA -3′ and WolPRANC792R 5′-TAACGCAAACGCTAAGCAAA -3′). Amplicons were successfully produced from singly infected *N. vitripennis* infected with *w*VitA and *w*VitB *Wolbachia* types as well as a *Gryllidae* sp infected with *Wolbachia* from the F supergroup (PNG0017). DNA was also amplified and sequenced from *Wolbachia* infected *Drosophila ananassae* and *Wolbachia*-free *D. ananassae* with an integrated *Wolbachia* chromosome in the nuclear genome

For a phylogenetic analysis, we used a subset of PRANC-domain sequences corresponding to proteins from *N. vitripennis*, the *Wolbachia* two endosymbionts of *N. vitripennis* (*w*VitA and *w*VitB), *D. ananassae* (*w*Ana) and *D. simulans* (*w*Sim) and six pox-viruses (Ectromelia virus, Molluscum contagiosum virus, Myxoma virus, Rabbit fibroma virus, Vaccinia virus and Yaba monkey tumor virus).

The original PRANC-domain definition from Pfam A covers 97 amino acids. To increase the sensitivity of the phylogenetic analysis, we extracted N- and C-terminal flanking regions of a length of up to 25 amino acids (depending on the number of residues C-terminal to PRANC) and created a multiple sequence alignment of these sequences with M-coffee (*S86*) This alignment is deposited in Treebase (www.treebase.org, ID SN4709). We then used this alignment with 1,000 bootstrap replicates for a Maximum Likelihood (ML) analysis with RAxML (*S87*). We estimated the most appropriate model settings using ProtTest (*S88*) (JTT substitution model (*S89*) with empirical amino acid frequencies and the gamma model of rate heterogeneity).

In order to assure that PRANC genes were actually components of the *Nasonia* genome and not viral or bacterial contaminant sequence, we used WU-BLAST (*S24*) to find all *N. vitripennis* trace sequences that matched PRANC genes (parameters W=100 E=e-20). We then took the mates of these reads and searched the mates against all non-PRANC genes to identify mate pairs in which one read matched a PRANC gene and one read matched a non-PRANC gene.

6. Screen of Evolutionary Rates of *Nasonia* vs. Honeybee Orthologs

Apis mellifera is the closest completely sequenced and annotated relative of *N. vitripennis*, yet their evolutionary distance prohibits dN/dS studies due to saturation. Thus we constructed maximum likelihood phylogenetic trees for over 3,000 orthologous groups with 1 *A. mellifera*, 1 *N. vitripennis*, and at least 1 gene from another species (either bodylouse, fly or beetle as an outgroup). A *Drosophila* entry was required for annotation of the orthologous group.

Orthologous group delineations were taken from (*S28*). Proteins alignments were constructed using MUSCLE (*S43*) and quality-filtered using Gblocks (*S90*). We calculated individual trees with phyML v2.4.4 (*S91*) using the JTT model of amino acid substitution and a gamma distribution over four rate categories. Trees were rooted with an outgroup and branch lengths extracted with scripts using bioperl. To compare branch lengths we calculated the rank difference with ranks calculated separately for each branch and submitted the resulting list to Gene Set Enrichment Analysis (GSEA) (*S92*). All statistical tests were corrected for multiple testing with Benjamini Hochberg's False Discovery Rate (FDR).

7. Gene Category Comparisons

a. Cuticular protein genes

Cuticular protein sequences with the R&R consensus (CPR proteins) were annotated using multiple comparisons to search both predicted genes and the raw genome data. All *Apis* and selected *Anopheles* CPR proteins were used in a TBLASTN search to identify similar *Nasonia* sequences. All *Nasonia* predicted proteins with the "chitin_bind_4" domain (corresponding to the R&R consensus region) were identified. Each existing annotation was checked to assure that it begin with an initiator methionine, had a signal peptide and a stop codon. When elements were missing they were searched for in nearby genomic regions and conventional splice sites were employed to recognize exons. Sequences were verified/corrected with available EST data and sequences from *N. giraulti*, and *N. longicornis*. In some cases, orthologs in additional species were used to assist in the annotation. Some predicted exons were eliminated if they coded for cysteine residues that are rarely found in mature CPR proteins, and some exons extended to reach a stop codon. Of the 62 CPR genes identified, eight were genes that had not been predicted, 19 others needed to be modified, eight to obtain signal peptides and the rest for other reasons.

b. Hexamerins in *Nasonia vitripennis*

The hemocyanin-like genes were identified in the official set of protein sequences (OGS v1.2) predicted from *Nasonia* genome assembly v1.0 using the program HMMER (*S93*) to search for the three hemocyanin domains available in Pfam database (*S46*); pfam03722: Hemocyanin_N; pfam00372: Hemocyanin_M; pfam03723: Hemocyanin_C). All *Nasonia* HEXs and PPOs were manually annotated in the wasp genome using Artemis software (*S94*). The nucleotide sequences of hemocyanin-like gene models were aligned against the EST database of *N. vitripennis* to check for expression evidence.

c. A comparison of oxytocin/vasopressin-like genes

Sequences corresponding to oxytocin and vasopressin-like preprohormone proteins were obtained from NCBI databases by searching directly for annotated genes or through TBLASTN and BLASTP searches (*S24*). Incomplete or unclear sequences were refined by FGENESH+ at http://linux1.softberry.com/berry.phtml (*S13*) using already known protein sequences as template. The nucleotide

sequences were aligned on their coded amino acid sequences with T-Coffee (*S35*) at http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffee_cgi/index.cgi through and edited in MEGA 4.0 (*S95*). The edited multiple nucleotides alignment was used for the phylogenetic analysis, which was based on Bayesian (MrBayes 3.1.2, (*S96*)) and parsimony (PAUP 4.0b20, (*S97*)) frameworks. The optimal model for sequence evolution for the Bayesian analysis, determined with ModelTest 3.7 (*S98*) under the AIC criterion, was GTR+I+G. Two parallel runs and four MCMC searches per run were performed for 5×10^6 generations. Burn-in, determined visually by plotting likelihood vs. generation, was set at 10^6 generations. Support for the parsimony tree was calculated by bootstrap analysis (1,000 replicates). Accession numbers of coding sequences are: *Eisenia fetida* Annetocin: AJ426471; *Platynereis dumerilii* Vasotocin: EF544399; *Strongylocentrotus purpuratus* Vasotocin: XM_001176026; *Daphnia pulex* Vasotocin: Not available; *Aplysia kurodai* Lys-conopressin: AB046867; *Lymnaea stagnalis* Preproconopressin: M86610; *Octopus vulgaris* Cephalotocin: AB108430; *Octopus vulgaris* Octopressin: AB056454; *Nasonia vitripennis* Inotocin: XM_001606497; *Tribolium castaneum* Inotocin: NM_001085362; *Danio rerio* Arginine vasopressin: NM_178293; *Catostomus commersonii* Vasotocin: M25144; *Catostomus commersonii* Isotocin 1: X16621; *Danio rerio* Oxytocin-like: NM_178291; *Platichthys flesus* Isotocin: AB036518; *Takifugu rubripes* Isotocin: U90880 AF468648 (locus with both genes); *Platichthys flesus* Vasotocin: AB036517; *Takifugu rubripes* Vasotocin: U90880 AF468648 (locus with both genes); *Plethodon shermani* Vasotocin: EF526214; *Bufo japonicus* Vasotocin: M16233; *Bufo japonicus* Mesotocin: M16232; *Typhlonectes natans* Arginine vasotocin: AF228336; *Taeniopygia guttata* Argininge vasopressin: XM_002190047; *Coturnix coturnix* Arginine vasotocin: AY786510; *Gallus gallus* Arginine vasopressin: NM_205185; *Ovis aries* Oxytocin: NM_001009801; *Bos Taurus* Oxytocin: NM_176855; *Homo sapiens* Oxytocin: NM_000915; *Mus musculus* Oxytocin: NM_011025; *Equus ferus caballus* Oxytocin: XM_001916032; *Homo sapiens* AVP: NM_000490; *Microtus ochrogaster* AVP: DQ269208; *Mus musculus* AVP: NM_009732; *Notomys alexis* AVP: AY856058; *Rattus norvegicus* AVP: NM_016992.

d. Meiosis: Phylogenomic inventory of meiotic genes

Protein sequences from *Apis* and other arthropods were used as queries in BLASTP and TBLASTN searches against the *Nasonia* genome to identify meiosis-related genes. For each gene present, nucleotide sequences were retrieved and protein sequences inferred. Amino acid alignments, including homologs from arthropods and other eukaryotes, were constructed using Clustal X (*S99*) and edited manually. Bayesian phylogenetic analyses were performed with MrBayes 3.1.2 (*S96*) and run for 10^6 generations with four Markov chains (one heated) using the WAG substitution model and eight gamma-distributed plus one invariable rate heterogeneity categories.

e. RNAi genes

Homologs of known small RNA genes were manually annotated from the RefSeq, Gnomon ab initio, and Fgenesh gene sets. A summary of these genes in shown in Table S4. *Nasonia* sequences were queried with protein sequences from known small RNA processing genes from *Drosophila melanogaster* and *Caenorhabditis elegans* (*S100, S101*). Therefore, homology statements are based on BLAST searches and protein sequence identity (e-value > 1e-30), and confirmed by reciprocal BLAST searches against the NCBI non-redundant database.

f. Odorant Binding Proteins

Genes encoding odorant binding proteins (OBPs) were located in the *Nasonia* genome using a combination of PSI-BLAST (*S102*) and HMMER (*S83*) searches. The PSI-BLAST PSSM and HMMER profiles were build using a collection of OBPs described in other insects, as described in (*S103*). Gene models were then manually constructed using the Apollo genome curation Software (*S22*), starting from the ab initio predictions if any was present and taking into account EST evidence when available.

VII. Evolutionary Genetics and Speciation

1. Evolutionary Rates Between *Nasonia* Species Methods

In order to examine evolutionary rates between species pairs (*N. vitripennis* vs *N. giraulti*, *N. vitripennis* vs *N. longicornis*, and *N. giraulti* vs *N. longicornis*), we aligned all RefSeq genes for each pairwise comparison using T-Coffee (*S35*). We then selected pairwise alignments which were in frame for their whole length, and had no premature stop codons in *N. giraulti* or *N. longicornis*. This set of alignments included 7,385 for *N. vitripennis* vs *N. giraulti*, 7,651 for *N. vitripennis* vs *N. longicornis*, and 7,367 for *N. giraulti* vs *N. longicornis*. We then estimated dN/dS for each gene pair using codeml in PAML (*S104*) with only one omega value across sites and branches. We subsequently screened 20 alignments of genes with high dN/dS values to ensure that no alignment errors had occurred. *Drosophila* Gene Ontology (GO; The Gene Ontology Consortium 2000) terms were then mapped to all gene pairs which were found to be 1:1 orthologs between *N. vitripennis* and *Drosophila melanogaster* in the comparative genomics analysis (see section VI.1). In order to test for accelerated evolution within certain GO terms, we randomly resampled dN/dS onto the gene pairs for each of the major GO term categories: process, function, and component, and calculated mean dN/dS values for each GO term 10,000 times to generate a null distribution for each GO term. We then compared the actual mean dN/dS of each GO term which appeared in at least 5 gene pairs to the null distribution to calculate a p-value. We corrected for multiple comparisons using a 5% false discovery rate with Q-value (*S105*). We further tested specific gene sets by

comparing dN/dS between *N. vitripennis* and *N. giraulti* for that set versus remaining RefSeq genes.

2. Analysis of Cytonuclear Incompatibility

Bulk Segregation Analysis of F2 Hybrid Males: To illustrate the effects of nuclear-mitochondrial incompatibility in species hybrids, we utilized the ability of the custom Nimblegen mapping array to quantify the amount of *N. vitripennis* vs *N. giraulti* DNA for each locus in bulk DNA samples. We pooled 100 surviving F2 haploid males from reciprocal crosses of *N. vitripennis* x *N. giraulti* with either a *N. vitripennis* or *N. giraulti* mitochondrion, and genotyped these bulk males using the microarray. DNA was prepared from bulk samples using a Puregene Gentra DNA extraction kit (Qiagen, Valencia, California, USA). This DNA was labeled and hybridized following NimbleGen's User's Guide: CGH Analysis v.3.0 methods with minor modifications. DNA was fragmented by either hydrodynamic shearing (HydroShear, GeneMachines) or sonication (Sonicator 4000, Misonix) and subsequently labeled by random primer labeling. Dual-color hybridization, post-hybridization washing and scanning were done according to manufacturer's instructions (NimbleGen's User's Guide: CGH Analysis v.3.0). Images were acquired using a GenePix Professional 4200A scanner with GenePix 6.0 software, and data from these arrays were extracted using the software NimbleScan 2.4 (Roche NimbleGen). In order to depict the amount of *N. vitripennis* vs *N. giraulti* DNA recovered on a genome-wide scale, we utilized the Illumina array genome map (*S38*). For each Illumina map marker, we calculated the average proportion of *N. vitripennis* vs *N. giraulti* DNA for all Nimblegen loci within 200 Kb of all Illumina loci which mapped to that specific marker. We then conducted a Wilcoxon matched-pairs signed ranks test (Perl script from http://www.fon.hum.uva.nl/rob/) for each Illumina marker to determine whether a significantly different proportion of *N. vitripennis* vs *N. giraulti* DNA was present in the two genetic backgrounds (*N. vitripennis* or *N. giraulti* mitochondrion).

3. Intraspecific Variation

To estimate the levels of intraspecific variation in all three *Nasonia* species, we sequenced 25 genetic regions for 5-19 strains from each species (Table S14) encompassing 10,463 bases of coding and 5,646 non-coding sequence. The primers and conditions of seven of the genes are given in (*S106*). DNA was extracted from one or more insects per strain using the DNAeasy kit (Qiagen). To clean the reactions before sequencing amplified reactions (8 μL) were incubated with 0.5 U shrimp alkaline phosphatase and 1.0 U of exonuclease I (Amersham, Piscataway, NJ) with the supplied buffer. Sequencing was performed directly from the amplified products using a BigDye v3.0 terminator sequencing kit and an ABI 3700 or 3730*xl* (Applied Biosystems, Foster City, CA) automated sequencer. The chromatograms generated were manually inspected and cleaned with Sequencher (Gene Code) and the sequences were aligned with Bioedit 7.0.1 (*S107*). For all of the genes that were resequenced for the three genome strains, no polymorphisms between these sequences and the genome sequences were observed. Additionally, no heterozygosities were observed in

any sequences. Analyses of the molecular genetic parameters were performed using DNAsp vs 4.10.2 (*S108*).

VIII. Biological Processes.

1. Sex determination

For RT-PCR and 5'- and 3'-RACE-PCR, one day old adult males, females, and gynandromorphs were studied separately. All RNA isolations were done using Trizol (Invitrogen). RNA concentration was determined spectrophotometrically and quality was checked on a denaturing agarose gel. 300 ng of total RNA was reverse transcribed with RevertAid H Minus First Strand cDNA Synthesis Kit using the poly-A adapter primer provided with the RLM-RACE kit. This template was subsequently used in a RT-PCR or 5'- or 3'-RACE-PCR. Both 5'- and 3'-RACE-PCR reactions were done using FirstChoice® RLM-RACE Kit (Ambion). All primers are listed in Table S42.

RT-PCR analysis was done on cDNA using 400 nM of primers NvTra_F2 and NvTra_R3. 1 µl of cDNA template was used in a standard Taq DNA polymerase (Roche) PCR with the following cycle settings: 94° C 5 min.; 94° C 30 sec.; 55° C 30 sec.; 72° C 45 sec. 40 cycles; 72° C 7 min.

3'-RACE analysis was done in two steps. First, 3 µl of cDNA was used in a PCR reaction with the 3'-RACE outer primer supplied with the RLM-RACE kit and 400 nM of the outer specific primer, NvTra_F1. 1 µl of this PCR product was then used in a nested PCR reaction with the 3'RACE inner primer supplied with the RLM-RACE kit and 400 nM of the inner specific primer, NvTra_F2.

For 5'-RACE 750 ng of total RNA from males and females separately was processed according to the manufacturers protocol. The nested PCR was done using 1 µl of cDNA in a PCR using Expand Long Template PCR System (Roche) with buffer 3 using 5' outer primer supplied with the RLM-RACE kit and 400 nM of NvTra_R3. The PCR cycle settings were as follows: 94° C 3 min; 94° C 35 sec.; 55° C 35 sec.; 68° C 5 min 35 cycles; 68° C 7 min. 1 µl of this reaction was used in a second reaction using 5' inner primer supplied with the RLM-RACE kit and 400 nM of NvTra_R2 or NvTra_R1 using the same PCR profile. The 3' and 5' RACE fragments were then cloned into pGEM-T vector (Promega) and Sanger-sequenced on an ABI 3130XL gene analyzer.

2. Diapause

Proteins were extracted from *Nasonia* larvae, digested, and analyzed by LC-MS/MS as previously described (*S109*). Using OMSSA 2.0.0 (*S110*), spectra obtained by mass spectrometry (linear ion trap) were matched against a database containing: RefSeq coding genes, RefSeq non-transcribed pseudogenes, GLEAN6 genes not overlapping with RefSeq, Gnomon genes not found in GLEAN6, and putatively contaminants including host proteins. Precursor and fragment tolerance were set to 0.8 Da, one missed tryptic cleavage was allowed, and the E-value cutoff was set at 0.01. The false discovery rate (FDR)

as determined by using a database with reverse sequences was 0 for accession numbers identified by ≥ 2 unique peptides and ≤ 0.95 % for accession numbers supported by 1 peptide (Table S43, available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html).

3. Venom proteins

Bioinformatic screening of the Nasonia genome: The NCBI RefSeq database was BLAST-searched against 387 amino acid sequences of known hymenopteran venom proteins, using both the stand-alone and the on-line version of the BLAST-software package. Based on an in-depth study of their conserved domain architecture and sequence identity, we retained a limited number of sequences for confirmative expression studies of the venom reservoir tissue by RT-PCR as outlined by (*S111*).

Venom sample preparation: Crude venom removed from the venom reservoir was desalted in a 5 kDa molecular weight cut-off centrifugal filter (Vivaspin). The final sample volume was reduced to 20 µl. The proteins were reduced and alkylated by respectively adding 2 µl of a 50 mM DTT solution in MQ-water for 10 min. at 100° C, and 2 µl 100 mM IAA solution in MQ-water for 2 hrs at 37° C. Proteins were digested overnight by adding 15 µl of a 0.1 µg/µl trypsin solution. The resulting peptide mixture was dried in a speedvac (Thermo Savant) and dissolved in 15 µl 5% ACN/0.1% formic acid (v/v).

LC-MS analysis: 10 µl digested venom extract was loaded for an off-line 2D-LC-MALDI-TOF MS experiment as previously described (*S112*). The tryptic peptides were first separated on a strong cation exchange (SCX) column, 150 mm × 0.3 mm, packed with POROS 10S (Dionex-LC Packings). The LC-effluent was fractionated in a multi well-plate at 5 min. intervals using a Probot spotting device (Dionex-LC Packings). SCX-fractions were dried and dissolved in 15 µl 5% ACN/0.1% formic acid (v/v). 10 µl of the SCX-fraction was used for LC-MALDI-TOF-MS, according to (*S112*). The peptides were separated onto a PepMap C18 analytical column, 150 mm × 75 µm (Dionex-LC Packings), at a 250 nl/min. flow rate. RP-LC effluent was spotted every 30 seconds on a MALDI-plate using the Probot spotter. Afterwards, 0.5 µl of a MALDI-matrix solution (10 mg/ml alfa-cyano-4-hydroxycinnamic acid dissolved in a 50% ACN/0.1% TFA/10 mM dibasic ammonium citrate) was spotted over the dried RP LC-fractions using the Freedom EVO robotic setup (Tecan).

MALDI-TOF/TOF measurements were performed on a 4700 Proteomic Analyzer (Applied Biosystems). MS-spectra were acquired by collecting 2000 sub-spectra at a laser intensity of 4300. A job-wide interpretation method, using all spectra from an LC run, selected automatically 5 precursor peaks per MALDI-spot for tandem MS. MS/MS-spectra were acquired by accumulating 5,000 sub-spectra at a laser intensity of 4,900. Finally MS/MS-spectra were subjected to a MASCOT (Matrix Science) database search using the GPS explorer v2.0 software (Applied Biosystems). Peptides were identified using a *N. vitripennis* Gnomon database.

For LC-ESI-FT MS analysis, 5 µl from all SCX-fractions described above were separated on an Agilent 1200 chromatographic system equipped with a Zorbax

300SB-C18 trapping column, 5 mm × 0.3 mm, and a Zorbax 300SB-C18 analytical column, 150 mm × 75 µm (Agilent). Samples were initially trapped via a capillary pump at a 4 µl/min flow rate, followed by separation on the analytical column at a 300 nl/min flow rate by applying a 35 min. linear gradient ranging from 2% ACN/0.1 formic acid to 80% ACN/0.1% formic acid in water. The LC-effluent was directly coupled to a Triversa NanoMate ESI source (Advion). The LTQ-FT Ultra mass spectrometer (Thermo Fisher Scientific) acquired MS-scans during the LC run at a 100,000 resolution. MS/MS fragmentation spectra were acquired in the LTQ XL Ion Trap for the three most intense ions in each MS spectrum. Raw LC-MS/MS data were subsequently analyzed with the Sequest database searching algorithm implemented in the Bioworks v3.3.1 software (Thermo Fisher Scientific). MS/MS data were searched against the Gnomon protein databases from *N. vitripennis*, concatenated with a shuffled decoy database generated with the Decoy Database Builder software (*S113*). According to the manufacturers guidelines, only peptide hits with XCorr values higher then 2.0, 2.5 and 3.5 for charge states +2, +3 and +4 respectively were retained in the final peptide list as a good criteria for positive identification. A summary of venom proteins identified using these methods is shown in Table S16.

4. Xenobiotics

Sequences encoding glutathione S-transferases (GST), cytochrome P450 (P450), and carboxyl/cholinesterases (CCE) were identified from the official set of gene (OGS v1.2) sequences predicted from the *N. vitripennis* genome project using the HMMER program (http://hmmer.janelia.org/) with protein domains for CCE (PF00135), GST (PF00043 and PF02798), and P450 (PF00067) described in the Pfam protein families database (*S46*). A significance (E-) value of at least 10^{-10} was used in the searches.

5. Pathogens, Symbionts, and Immunity

a. Annotation of immunity genes

Immune repertoire analysis: A global scan to identify putative immune-related genes in *Nasonia* was based on a library of Hidden Markov Model (HMM) profiles built using carefully selected groups of immune-related proteins. The first set was built from groups of manually curated protein sequences of immune-related genes from three Dipteran species – *Drosophila melanogaster*, *Anopheles gambiae*, and *Aedes aegypti* - based on the Waterhouse *et al*. 2007 (*S114*) immunity analysis with all data accessible from the ImmunoDB resource: http://cegg.unige.ch/Insecta/immunodb. The second set of HMMs was built from automatically defined orthologous groups (*S28*) across five vertebrate and five insect species, using *Aedes* gene identifiers as seeds to identify groups belonging to the immune repertoire. All selected groups of sequences were aligned using MUSCLE (*S43*) and the HMMs were built and calibrated using HMMER (*S115*). The HMM library therefore includes manually and automatically defined immune-related genes and gene (sub)families of anti-microbial peptides,

gram-negative binding proteins, caspases, catalases, clip-domain serine proteases, c-type lectins, fibrinogen-related proteins, galactoside-binding lectins, inhibitors of apoptosis, IMD pathway members, JAKSTAT pathway members, lysozymes, MD2-like receptors, peptidoglycan recognition proteins, peroxidases, prophenoloxidases, scavenger receptors, superoxide dismutases, spaetzle-like proteins, serpins, thio-ester containing proteins, toll-like receptors, and TOLL pathway members. These HMMs were then run against the five different available *Nasonia* gene prediction sets: GLEAN, RefSeq, Augustus, Gnomon, and Fgenesh to scan for immune-related genes. They were also run against *Anopheles* (AgamP3.4), *Aedes* (AaegL1.1), and *Apis* (preRelease2OGS+*Abinitios*): cross-referencing with the 'known' immune repertoires defined in (*Anopheles* & *Aedes*) (*S114*) and (*Apis*) (*S116*) thereby provided a cross-check on the HMM performances in terms of false positives (non immune-related genes falsely identified as such by the HMM scan) and false negatives (true immune-related genes missed by the HMM scan). Importantly, these scans were performed against the five available predicted *Nasonia* proteomes – if any given gene is not represented in any of these prediction sets it cannot be picked up by the scan, or if the gene model is truncated/incomplete it will likely score poorly and therefore not pass the required scoring thresholds. The scan may over-predict the group of Clip-domain serine proteases due to the abundance of proteases, only some of which possess the defining clip-domain, but manual searches for the Clip domain on the set of genes produced by the scan identified the true matches.

Further identification of predicted proteins in the major immune pathways (Imd, Toll, and Jak/STAT) and of particular problematic gene families and classes (CLIP serine proteases, antimicrobial peptides) was carried out using alignments to specific predicted domains, and both single-course and iterative (PSIBLAST) alignment algorithms. Best fit models to the proposed Dorsal and PGRP transcripts were constructed on the basis of EST evidence (in the case of a large PGRP-LC-like locus with multiple exon candidates) and on alignments to *Apis* gene models.

b. Expression of PGRP/LC

To study host gene expression in response to *Wolbachia* infection, *Nasonia girault*i males and females carrying extreme bacterial loads (strain IntG[12.1]) and tetracycline-cured controls (strain IntG[12.1]T) were profiled for expression differences in eleven immunity genes: beta-1,3-glucan recognition protein, c-*Jun* protein, thioester-containing protein *TEP III*, *Dredd*, *Relish*, two domains of the *PGRP-LC*-like locus, and Toll pathway genes *MyD88*, *dorsal*, *cactus*, *tube*, and *Spatzle 1B*. RNA was harvested from adults infected and uninfected with *W. pipientis* using Trizol (Invitrogen). Following cDNA synthesis using the ThermoScript kit (Invitrogen) and a 1/50 dilution of the template, transcript levels were measured by RT-PCR in duplicate and analyzed by the Pfaffl method for relative gene expression (*S117*).

6. Neurohormones and ion channels

We searched the *N. vitripennis* genome sequence assembly v1.0 and the unassembled contigs with protein sequences corresponding to various insect and mammalian neuropeptide precursors and neurohormone GPCRs, using TBLASTN at the Baylor College of Medicine (BCM) blast server (http://blast.hgsc.bcm.tmc.edu/blast.hgsc?organism=9). The regions that were hit with highest scores were checked for corresponding gene models in the GENBOREE set (http://www.genbore.org/java-bin/login.jsp) that contains the automated GLEAN2 predictions incorporating the results from multiple gene prediction programs and by manual corrections. Subsequently, the gene sequences were manually curated and annotated on the BCM *Nasonia* annotation server.

Signal peptides were predicted by the SIGNALP server (http://www.cbs.dtu.dk/services/SignalP/) transmembrane regions were predicted by the TMHMM server (http://www.cbs.dtu.dk/services/TMHMM/). Multiple sequence alignments were performed with ClustalW (http://www.ebi.ac.uk/clustalw/) or using the Lasergene software package (DNASTAR) with manual adjustments. Phylogenetic analyses and bootstrapping were done using Phylip protdist (http://mobyle.pasteur.fr/cgi-bin/portal.py?form=protdist). Tree diagrams were drawn using the phylogenetic tree printer Phylodendron (http://iubio.bio.indiana.edu/treeapp/treeprint-form.html).

7. Courtship

The *Wolbachia*-cured, inbred lines IV7R2 (*N. longicornis* – paternal line) and RV2x (*N. giraulti* – maternal line) were used to produce F_1 hybrid females. These females were isolated as virgin and provided individually with host-fly pupae to produce all-male F2 offspring.

Wasps were bred at 20-25°C in incubators under constant light. Test animals were of controlled age (2-3 days post-emergence). To obtain standardized material males were isolated in glass tubes at least 12 hours prior to testing. Males were virgin whereas females were already mated to prevent premature termination of courtship due to copulation. The recorded components of male courtship behavior have been shown to be unaffected by the mating status of the female, except the total number of series, since *Nasonia* males are incapable of distinguishing between mated and virgin females (*S118*).

Nasonia male courtship is characterized by periodically repeated series of motor patterns (*S118, S119*). Successive headnod series are separated by pauses. An interval starts with a head nod series. The interval between the 1st head nods of two consecutive series is termed a cycle. We recorded in chronological order (1) time of the males' rapprochement toward the female called latency (interval between introduction of the female and moment of mounting), (2) the number of headnods in each cycle for the first four cycle (number of headnods) and (3) the duration of cycle time also for the first four cycles (cycle1-4), and (4) the total number of cycles until dismounting of the male (total number of cycles). Observations started with the introduction of a female

into the tube and ended with either the male dismounting after a bout of courtship, or after 10 minutes if a male did not mount the female. A total of 121 males were phenotyped for these 10 male courtship components.

All males were genotyped for 29 microsatellite markers at an average marker distance of 20 cM distance. Markers are named after their scaffold and their position within this scaffold (e.g. Scaf1_3410194 = marker on scaffold 1 at position 3410194).

MapQTL 4.0 (*S120*) was used to perform an interval and MQM (marker-QTL-marker). Significance thresholds for each individual trait were determined using an implemented permutation test.

SOM Supporting Text

I. Validation of Gene Sets

1. Yellow and Royal Jelly-like proteins

Yellow-like proteins are products of an intriguing gene family that have been found only in insects and in certain bacterial and fungal species (*S121, S122*). While their functional significance in microbial organisms remains unknown, in insects they have been implicated in diverse functions related to pigmentation, development and sexual maturation (*S122*). Multiple genes encoding Yellow-like proteins, ranging in numbers from eight in *D. pseudoobscura* to 20 in *A. mellifera,* have been uncovered in each of the insect genomes characterized thus far. Phylogenetic analyses of these genes revealed functional specialization occurring during evolution of Yellow proteins. The genomic architecture of the *yellow* gene family in the honey bee suggests that in ancestral bees subsequent duplications of one member of the *yellow* cluster produced a new gene subfamily (*royal jelly*), which evolved new functions relevant to social behavior (*S122*). These proteins are major components of the queen bee larval food and are typically referred to as Major Royal Jelly Proteins (MRJPs). In the bee genome, MRJPs are encoded by a tandem array of nine active genes (and one pseudogene) that are clustered inside the chromosomal region harboring members of the *yellow* gene family (*S122*).

The *Nasonia* genome contains 26 genes encoding Yellow-like proteins including 10 similar to *mrjp*s in honey bees (Table S44). Five *Nasonia rjpls* (1-5) form a tandemly arranged gene cluster flanked by *yellow* genes on the scaffold 42 (GenBank NW_001818237.1). This genomic region is syntenic with the honeybee *mrjp* gene cluster. This contig also harbors one additional *rjpl* in a distal part. Three genes (*RJPL6-8*) are tandemly arrayed on the scaffold 143 (GenBank NW_001815059_1) and one *rjpl* is located on the scaffold 1011 (NW_001814584) (Fig. S17). Interestingly, these RJPL proteins in *Nasonia* represent a new gene subfamily rather than true orthologs of *mrjp*s in *Apis*. This notion is supported by several lines of evidence: i) phylogeny of Yellow-like proteins shows two independent RJPL and MRJP branches supported by high bootstrap values suggesting an independent evolution of these proteins in *Apis* and *Nasonia* (Fig. S5), ii) none of *rjpl* genes contains “intron 4” found in all *Apis mrjp*s*,* iii) none of *Nasonia* RJPL proteins contains C-terminal repeats characteristic of most *Apis* MRJPs, iv) all *Nasonia rjpl* genes have the same intron structure and transcriptional phases suggesting a common origin, v) *mrjp*s are highly expressed in honeybees, whereas only several ESTs exist for three out of ten *rjpl*s.

Yellow-e3 gene identified previously as an originator of *mrjp* genes in honeybees has a clear homologue in *Nasonia*. Its intron structure differs from that of MRJPs only by the absence of the third intron that was gained early in the evolution of the MRJP group. As mentioned above *Nasonia rjpl*s lack the forth

intron but contain the third intron suggesting that they evolved after its gain. Taken together, MRJPs and RJPLs genes evolved from a common originator, but multiple gene duplication events occurred in the bee and wasp lineages independently resulting in two related but functionally diverse protein families. MRJPs developed nutritional function as components of the larval food; the function of RJPLs remains unknown. Tiling array expression data shows that Yellow-like proteins have diverse expression patterns across developmental stages and tissues (Table S7).

3. Cuticular Protein Genes

A common feature of the majority of cuticular proteins from diverse arthropods is the presence of an extended version of the Rebers and Riddiford (R&R) Consensus, a conserved domain of about 63 amino acids that binds to chitin (*S123*). Genes that code for proteins with this consensus are represented by large families in dipterans. There are 156 such genes (called *CPR*) in *Anopheles gambiae*, about 1% of all predicted genes (*S124*). The number in other mosquito species is even higher (*S125*) and *Drosophila melanogaster* has 101 (*S126*). In the silkmoth, *Bombyx mori*, 148 *CPR* genes have been identified (*S127*). Over 90% of the predicted *An. gambiae* CPR proteins have been identified by mass spectrometry in preparations of cuticle (*S128*). Hence, related sequences in other species are apt to be authentic cuticular proteins and not figments of annotation.

Apis mellifera has only 32 genes that code for CPR proteins (*S67*). Another cuticle protein family, apidermins, was recently identified in *Apis* with three members (*S129*). Other families of cuticular proteins are known but the numbers in any one species are generally small (*S130*), although individual families may have expanded in particular lineages such as the 27 Tweedle genes in *D. melanogaster* (*S131)*. The protected and provisioned environment in which *Apis* resides until adulthood may account for the paucity of CPR genes. Alternatively this may be a hymenopteran characteristic. Thus it was of interest to learn the number of genes coding for cuticular proteins in *Nasonia*, although as a parasitoid it too may have relatively unspecialized larval and pupal cuticles, and hence a lower number of cuticular protein genes. *Nasonia* appears to have 62 CPR genes. Hence while the numbers of CPR genes in *Nasonia* exceed those in *Apis*, they, and genes coding for cuticular proteins in other cuticular protein families, are still lower than in the other species where complete genomic sequences are available.

Twenty-four of the *Apis* CPR proteins (75%) had convincing orthologs in *Nasonia* (see Table S11). All others were represented by two or more closely related *Nasonia* sequences. All *Nasonia* CPR sequences had clear orthologs or paralogs in *Drosophila*. While the extended R&R Consensus region is, on average, 81% identical among *Apis*/*Nasonia* orthologs, the entire protein is far less conserved with an average of 58%. The presence of clear orthologs across orders and about 300 myr suggests that particular proteins are serving important and distinct (as yet unknown) functions.

4. Olfactory Binding Proteins and Chemosensory proteins

Odorant binding proteins (OBPs) and chemosensory proteins (CSPs) are two gene families that are believed to transport odorant molecules through the sensillum lymph to the olfactory receptors on the membrane of chemosensory neurons. Several members of these families have, however, been found in non olfactory tissues.

The *Nasonia* genome encodes 90 OBP genes, including 8 pseudogenes (Table S17). This contrasts with 21 honeybee OBPs. Two highly conserved pairs of OBP orthologs were found between *Nasonia* and the honeybee: Nvit76/AmelOBP1 (53% amino acid identity, 76% similarity) and NvitOBP2/AmelOBP10 (79% amino acid identity, 86% similarity). AmelOBP1 is specifically expressed in the antennae and has been suggested to bind to a major component of the queen pheromone (*S132*). AmelOBP10, however, is only expressed in the brain, reaching high levels in late pupae and newly emerged adults (*S103*).

Most of the other *Nasonia* OBPs are grouped in a number of lineage specific expansions, typically co-located in genomic clusters. One of these groups is made of genes with two concatenated OBP domains, but is not orthologous to the groups of *Drosophila* OBPs containing two domains.

The Nasonia genome encodes ten CSPs. Four of the six honeybee CSPs form orthologous pairs with *Nasonia* genes. These include AmelCSP5 (NvitCSP10), also known as *uth* (unable to hatch), involved in embryonic development (*S133*).

II. Genome Annotation

1. de novo Repeat Library Construction

The estimated complete genome size of *Nasonia vitripennis* is 332.5 Mb (*S134*), but only 295 Mb are represented in the genome assembly. The 37.5Mb of sequence that is unaccounted for probably represents repetitive centric and telomeric heterochromatin. In addition there were 56Mb of N's in the version 1 of the *Nasonia* genome assembly, which presumably represents a baseline for the repetitive fraction from the assembled genome regions (19%). Our measurement of the repetitive fraction within the 239Mb of non-N sequence in the *N. vitripennis* genome assembly was initially limited by the lack of species-specific repeats identified in hymenopterans. The closest annotated relative, *Apis mellifera*, has only a few R2 and mariner type transposons (*S67*). Since the use of repeat libraries from even phylogenetically closely related species underestimate the repeat composition of the genome under study, we undertook a *de novo* transposable element analysis using the PILER-DF program (*S42*) similar to analyses performed for Dipteran species (*S49*). Briefly, we self-aligned the complete *N. vitripennis* Release 1 genome and extracted regions that were duplicated three or more times, which we defined as the signature of transposable element (TE) sequences (Table S43 available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html).

We found 6,342 elements repeated three or more times in the genome assembly, which corresponded to 1,195 unique PILER-DF predictions that were less than 90% identical over 90% of their length. Of these nine were deemed to be false positives and identified various histone sequences (Table S29 available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html) On average, PILER-DF predictions in *N. vitripennis* were 173bp (median 295bp), with the shortest and longest elements measuring 101bp and 2,135bp respectively. Therefore, most predictions are significantly shorter than well-described, functional TEs in other model systems. For example, the annotation of retroid sequences using the Genome Parsing Suite (GPS) detection system (see Material and Methods, (*S50*) generated 6523 retroid annotations with significantly longer average length (1486bp) (Table S30 available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html). We classified PILER-DF predictions with RepeatMasker (*S44*) and BLAST (*S24*) using alignments scoring greater than 50 bits to known transposable element motifs including integrase, reverse transcriptase, transposase, and other domains downloaded from Interproscan (*S135*). Of the 1,195 PILER-DF predictions, 160 (13.3%) had similarity to either long terminal repeat retrotransposons (LTR, 93 families), long interspersed nuclear element (LINE)-like retroposons, 59 families), or both (8 families) types of retroid elements. These repeats were mainly R1 (52), Copia (20), and Gypsy (19) retrotransposons, while 48 (4%) of the PILER-DF predictions had similarity to mainly mariner (10), transib (6), and newly described Chapaev (7) DNA transposons (Repbase Reports 8(2), 59-59 (2008)). We observed 8 PILER-DF predictions that had similarity to both DNA transposons and retrotransposons. These may represent composite repeat nests that have undergone recent segmental duplication or could reflect artifactual alignments due to misassemblies. More in-depth analysis of retroid elements using the GPS system yielded more specific annotation of retroids and identified 75 distinct LTR and LINE-like families in *N. vitripennis* (Table S45 available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html). Nasonia shows an unusually large diversity of retroids including rDNA specific retroelements R1 and R2 that span the entire range of diversity found in arthropods (*S136*).

Many of the PILER-DF predictions (231, 19%) appear to represent simple repeats (61, 3%), low complexity sequence (126, 11%), or satellites (14, 1%). Strikingly, 786 (66%) of the PILER-DF predictions could not be identified by RepeatMasker or BLAST alignment to known TE proteins. We classify these as novel transposable elements simply by the fact that they occur at least three times in the genome. However, it should be noted that several of these elements may represent low complexity or simple repeat sequences and these predictions do not represent full-length TEs.

2. Gene and Genome Masking

While standard RepeatMasker using the 79-element Repbase library for *Nasonia* identifies 24% (73Mb) of the genome assembly as repetitive (including 56Mb of N's), addition of the PILER-DF libraries identified an additional 21 Mb of the genome as repetitive (94Mb total). The sequenced and assembled portion of the genome has 324,694 repeat regions identified that span 49Mb, with many more interspersed (57%) and non-interspersed repeats (43%, Table S2 and Tables S31 and S32 available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html) if novel PILER-DF predictions are factored in. LINE and LTR retrotransposons outnumber DNA TEs by nearly 10-fold and putative novel TEs represent 33% of the total repetitive regions annotated. The diversity of identified LTR retrotransposons (77 families, 42 novel), LINE-like retroposons (5 families, 0 novel), and DNA transposons (22 families, 16 novel) is comparable to other insect genomes. About 10% of sequences identified were simple repeats and 16% of the genome was identified as satellites (Table S2 and Table S34 available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html).

We also undertook a BLAST analysis to map nine known repeat elements that have been shown to be localized on B chromosomes (Psr105-3, Psr18-1, Psr22-2, nv79-16a, nv85-7, nv126-6, nv104-6, and AAAGTCT[T/C]GACTT) (Table33 and Table S34 available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html) Psr repeats are found exclusively on *N. vitripennis* paternal sex ratio (PSR) B chromosomes, while some of the nv- repeats also found on autosomes and in related species (*S47*). In total, we found 941 unique scaffolds that contained these 11,780 repeats with 80-100% sequence identity, representing 800,119bp of the genome assembly (0.3%). Notably, the psr-type repeats were usually fragmented and only covered 8-24% of the canonical repeat length, while full-length nv-type repeats could readily found. We have not searched the reads not utilized for the main assembly for these repeats, but it is highly likely that the remaining unsequenced portion of the *Nasonia* genome is composed of these and other repeat sequences.

One problem with automated genome annotations in newly sequenced genomes is that the lack of repeat libraries prevents genome sequence from being masked properly, giving rise to gene predictions contaminated with exons from transposable elements. Such features are erroneously included in subsequent analyses and may be printed on microarrays and other genomic resources. Identifying genes with such exons is important to the accuracy of further genome-wide analyses. We found that 406 RefSeq genes (4.3%), 906 GNOMON predictions (24.3%), and 1538 GLEANR consensus genes (24.6%) had similarity to our repeat library when scanned with RepeatMasker. Genes with similarities to repetitive regions are listed in Table S34 available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.

III. Gene Expression

1. Epigenetic Control

In insects, epigenetic control of gene expression has only been studied in *Drosophila melanogaster;* however, the recent genome sequencing of various insect species (*Tribolium castaneum, Apis mellifera,* and *Nasonia vitripennis)* allows us to examine the phylogenetic distribution of genes associated with epigenetic control, specifically Polycomb Group (PcG) and Trithorax Group (TrxG) proteins, and other histone modifications. We found that the genes associated with epigenetic control are relatively well conserved between these insects. However, some differences are shown in Table S46. The copy numbers of *Extra sex comb, Pleiohomeotic, JMJD,* and *HP1* genes are specifically increased in *Drosophila*, and *Zeste* is only found in *Drosophila.* No difference is found between *Apis* and *Nasonia*.

2. DNA methyltransferase

Annotation of DNA methyltransferases: Four DNA methyltransferase (*Dnmt*) orthologs were identified in the *Nasonia* RefSeq gene set. Based on sequence similarity (described below) to the C-terminal catalytic domains of the mammalian *Dnmt* genes, three genes (*NvDnmt1a, 1b* and *1c*) were assigned to subfamily 1 and one (*NvDnmt3*) to subfamily 3, which code respectively for the maintenance and the *de novo* DNA methyltransferases (*S60*). N-terminal domains characteristic of mammalian DNMTs are present in the *Nasonia* enzymes and are consistent with the subfamily assignment. A *Nasonia* ortholog of *Trdmt1* (tRNA aspartic acid methyltransferase 1) formerly known as *Dnmt2* was identified but not characterized further (*S137*).

Confirmation of gene expression and gene models: Expression in various developmental stages and the exon-intron structure of the RefSeq gene models was confirmed for *NvDnmt3* and each of the *NvDnmt1* genes by RT-PCR using primer pairs that flank predicted exon-intron boundaries. For the *NvDnmt1* genes, in some cases a PCR product was cloned and sequenced to confirm the splice boundary but, typically, the gene model splice sites were assumed to be correct if the RT-PCR products were of the predicted size. An example of these experiments is shown in Figure S16. For the *NvDmnt3* gene, dideoxy sequencing of RT-PCR products spanning exons 2 through 15 (out of 18 exons) of the gene model directly confirmed the predicted sequence for the most abundant transcript.

Pairwise comparison of the polypeptides encoded by the wasp, honeybee and human Dnmt genes: The predicted NvDMNT1A and 1B proteins are more similar to each other (62% amino acid identity and 78% similarity along the entire length of the genes) than either is to NvDNMT1C (44% and 61%, and 41 and 59%, respectively). NvDNMT1A is the most similar to *Apis* DNMT1A and 1B with ~62% identity and 76% similarity, and NvDNMT1C is the least similar with about 39% identity and 58% similarity to AmDNMT1A and 1B. Likewise NvDNMT1A is highly similar to human DNMT1 (with 52% identity and 68% similarity) along the length

of the gene while NvDNMT1C is the least similar of the three with 37% identity and 54% similarity to HsDNMT1.

The NvDNMT3 protein is longer than the honeybee and human orthologs and in all cases, the N-terminus of the Nasonia ortholog failed to align with either the honeybee or human proteins. However, NvDNMT3 shows 43% identity and 63% similarity to AmDNMT3 along the entire length of the honeybee protein and, excluding the N-terminal amino acids, about 33% identity and 52% similarity to both HsDNMT3A and HsDNMT3B.

C-terminal methyltransferase domain: Cytosine-5 DNA methyltransferases in prokaryotes and eukaryotes share a similar catalytic mechanism reflected in the conserved sequence motifs found in these enzymes across distantly related taxa (*S60, S138, S139*). Pairwise comparison of the C-terminal catalytic domains encoded by the Nasonia *Dnmt* genes with honeybee and human genes shows striking sequence similarity. NvDNMT1A shows about 78% identity and 88% similarity to both AmDNMT1A and 1B and 70% identity and 82% similarity to the HsDNMT1 (NvDNMT1B shows similar trends with slightly less homology than NvDNMT1A to the honeybee and human proteins). Interestingly, the NvDNMT1C methyltransferase domain shows about the same identity (48-50%) and similarity (65-68%) to the honeybee and human DNMT1 catalytic domains as it does to the *Nasonia* DNMT1A and 1B domains. Figure S15A shows a phylogram of the wasp, honeybee and human DNMT sequences illustrating the divergence of NvDNMT1C from NvDNMT1A and 1B.

Like the *NvDnmt1* genes, *NvDnmt3* encodes a conserved methyltransferase domain that exhibits 59% identity and 75% similarity to the AmDNMT3 and 44-45% identity and 63-65% similarity to the human DNMT3A and 3B proteins.

The catalytic domain of C-5 methyltransferases is organized into conserved sequence motifs. The most highly conserved motifs are found in the predicted amino acids sequences encoded by each of the *NvDnmt* genes (Fig. S15B, C). Motifs I and X contain the binding site for the cofactor S-adenosylmethionine (AdoMet) and motif IV includes the cysteine nucleophile required for catalysis (*S60*).

N-terminal domain structure: Although the *Dnmt* subfamilies share a similar C-terminal catalytic domain, they differ with respect to conserved N-terminal domains that play significant roles in the protein-protein interactions and, along with the C-terminal catalytic domain, DNA substrate specificity (*S60, S139*). The N-terminal domain architecture of the NvDNMT proteins is consistent with the subfamily assignments. Like the methyltransferases encoded by mammalian *Dnmt1* genes, all three NvDNMT1 proteins contain two BAH (bromo-adjacent homology) domains and a zf-CXXC (CXXC zinc finger) domain. Likewise, NvDNMT3 contains the PWWP characteristic of mammalian DNMT3.

Phylogenetics: Figure 1 shows phylogenetic relationships of *Nasonia* and DNA Methylation Toolkit. The human Dnmt3 subfamily members include Dnmt3A, 3B and 3L. The catalytically inactive DNMT3L protein is required for imprinting in mammals and has not been found in non-mammalian species (*S140*). A second, highly divergent copy of Dnmt3 is present in both *Acyrthosiphon pisum* and

Daphnia pulex. In *Pediculus humanus* and *Ixodes scapularis* one Dnmt1 gene copy consists of a single exon ORF and is divergent in sequence.

3. Genome-wide analysis of CpG content

Expected to observed CpG content ("normalized CpG") is an indicator of DNA methylation level, reflecting mutational biases of methylated CpG to TpG (*S101*). Vertebrate genomes show reduced normalized CpG except within 'CpG islands'. In *Apis*, a bimodal distribution of normalized CpG is found in coding exons, suggesting the presence of hyper- and hypo- methylated genes (*S141*). Likewise, the distribution of normalized CpGs in *Nasonia* strongly indicates DNA methylation (Fig. S2); *Nasonia* introns exhibit a clear distinction between low and high CpG regions, and exons show a broad distribution not found in insects without CpG methylation (Fig. S2).

4. Differing Methylation Profiles of *N. vitripennis* genes

Figure S3 shows the results of methylation detection of the following five RefSeq genes: XM_001607338.1 (vitellogenin), XM_001600728.1 (epithelial membrane protein), XM_001600593.1 (polypeptide of 976 amino acids), XM_001606530.1 (eIF 2a kinase), and XM_001601041.1 (eIF2B-gamma protein). The first two genes have relatively high CpG O/E (1.002 and 1.579) while the other three genes have low CpG O/E (0.459, 0.420, and 0.362, respectively). It has been shown that in human and sea squirt, regions with high CpG O/E tend to be hypo-methylated while those with low CpG O/E are likely to be hyper-methylated (*S81, S142*). In *N. vitripennis*, we observe a comparable phenomenon: the first two genes harbor relatively few methylated cytosines while most examined CpGs in the last two genes are methylated. This suggests that CpG O/E may be used as an indirect indicator of the level of DNA methylation in *N. vitripennis* genome.

IV. Mapping of Scaffolds & Genotyping

1. Visible marker and centromere locations

Decades of classical mutational genetic studies in *Nasonia* have resulted in a genetic map based on visible mutant markers (*S143*). We have mapped the approximate location of a number of these markers in the genome sequence (Fig. 2) by introgression and recombination mapping between *N. vitripennis* and *N. giraulti,* using a competitive genotyping hybridization microarray and PCR.

The location of two *Nasonia* centromeres can be inferred from the visible marker map. Specifically, centric fragments (armless chromosome fragments containing centromeres) have been produced that contain the wildtype alleles of *or-123* and the *R*-locus gene *st-DR,* as demonstrated by complementation tests (*S144-S146*). These chromosome fragments are cytologically very small with respect to the other chromosomes, suggesting that *or-123* and the *R*-locus are

physically located in or near centromeres. We mapped the genomic location of *or-123* and the *R*-locus using recombinants from between-species introgression lines. Both *or-123* and the *R*-locus map near regions of low recombination and gene density (Fig. 2), supporting the hypothesis that these low recombination regions represent the centromeres.

V. Genome Feature Analyses

1. Heterochromatin

Given the extremely low percentage of the *A. mellifera* genome that was identified as repetitive (26/260Mb genome) (*S67*) and scant number of characterized TEs in hymenopterans, it is notable that the *Nasonia* genome assembly has such a high repeat content (33% overall). *Nasonia*'s repeat content appears to be comparable to Dipteran genomes, where 25-60% of the genome assembly is comprised of repeat-rich regions (*S49*)). However, when the N and unsequenced regions are added to the presumed repeats from the non-assembled portion of the genome, *Nasonia* appears to have an overall repeat content of ~40% (94Mb+37.5Mb/333Mb), making it one of the most repeat-rich insect genomes studied. One intriguing possibility is that the elevated repeat content of the *Nasonia* genome is higher because of its parasitoid life history. That is, it is possible that transposable elements from host species have horizontally transferred into the *Nasonia* genome.

We were interested in the distribution of repeats across the genome and we fragmented the 6,181 scaffolds in the *Nasonia* genome assembly into 8,459 fragments that were 100kb or smaller pieces. We then calculated the total repeat content of those scaffolds and plotted their frequency distribution (Fig. S1A, S1B). In most metazoan species the genome is organized into euchromatin, with relatively little repetitive sequence and repeat-rich heterochromatin mainly found at the centromeres and telomeres. Based on the observed difference in frequency of high and low repeat content scaffolds, we defined any region greater than 16% repeat content in *Nasonia* to be heterochromatic, while regions less than 16% are defined as euchromatic (Fig. S1B). We investigated the spatial pattern of repeats in the *Nasonia* genome (Fig. S1) and defined regions of greater than 16% repeat content as putative heterochromatic sequence (4,919 segments; 127Mb), while regions less than 16% were defined as euchromatic (3,540 segments; 167.8Mb). When these putative heterochromatin regions are added to the presumed repeats from the non-assembled portion of the genome (37.5Mb), the heterochromatic regions in *Nasonia* appears to span 49% of the genome (37.5 Mb unsequenced + 127Mb of >16% repeat content scaffolds), making *Nasonia* heterochromatin larger and more repeat-rich than most insect genomes studied so far.

Given the large predicted size of heterochromatin in *Nasonia*, we were interested in the genes residing in that region. Recent studies suggest that many genes can be repositioned in the genome, moving from euchromatin to heterochromatin over evolutionary times (*S147*). Overall, 90% of the RefSeq

genes (Fig. S1C) are present in regions with less than 16% repeat content (euchromatic regions) and 10% of the RefSeq genes are located in putative heterochromatic high repeat content regions. Of ~150 RefSeqs that were hand annotated in the highest repeat content regions, 73 turned out to be transposable elements. While roughly half of these RefSeq-supported genes appear to be TEs, when we compared the distribution of ab initio predicted genes, nearly half resided in the high repeat content scaffolds (Fig. S1C). Since the ab initio repeat finders were run before the enhanced repeat analysis described, it is likely that the majority of these ab initio results represent TE ORFs that are misidentified as genes. Nonetheless, recent evidence from *D. melanogaster* suggested that repeat-rich centric heterochromatin contains many more expressed genes than previously expected (*S53*). Therefore, we identified putative orthologs to the *D. melanogaster* EST-supported heterochromatic genes using TBLASTN and used existing RefSeq and other evidence to annotate the complete gene models. Strikingly, only one of the 150 cDNA-supported *D. melanogaster* heterochromatic genes tested appeared in similar repeat-rich regions in *Nasonia*. These results suggest that genes are capable of adapting to significantly different chromatin complexes over evolutionary time.

We were also interested in potential heterochromatin genes that were in high-repeat content regions in *Nasonia*. We identified ~1000 well-supported genes in the *Nasonia* scaffolds with greater than 16% repeat content. Indeed, assuming that half of the 567 heterochromatic *Nasonia* RefSeq genes are in fact, TEs, that would leave ~280 potential genes in high repeat content regions, comparable to the number observed in *D. melanogaster* (*S49*). There were no obvious patterns for types of genes or gene families that were embedded in repeat-rich regions when analyzed with GO term finder (*S148*).

3. GC Content

Animal genomes are not uniform in their long-range sequence composition, but are composed of a mosaic of sequence stretches of variable lengths that differ widely in their GC compositions. Whether these stretches meet the criteria of isochores (sensu (*S149*)), or should better be referred to as GC-content domains (*S150*) is a matter of debate (*S80, S151-S153*). In all animals studied so far, the distribution of GC-content domain lengths (plotted on a log-log scale) was found to follow a heavy-tail distribution with power-law decay exponents ranging from –1.5 to –2.5 (e.g., (*S80*). The genome of the *Nasonia* is no exception in this respect. In other words, the compositionally homogeneous segments in its genome - as in all other genomes studied so far - do not have a characteristic length; rather, there is an abundance of short segments and only a small number of long ones. On average, the long GC-content domains in *Nasonia* and honeybee are shorter than in the two dipterans, which in turn are shorter than those in vertebrates. A comparison of the distributions of GC-content lengths among *Nasonia vitripennis*, *Apis mellifera*, *Tribolium castaneum*, *Anopheles gambiae*, *and Drosophila melanogaster* genomes is shown in Figure S18. Interestingly, *Nasonia* has the highest abundance of small size GC-content domains (3 kb - 17 kb) relative to the other insect genomes and its GC content

spans from 18% to 72%. Only a small fraction of the homogeneous domains are longer than 300 kb however their mean GC content (45.3%) differ significantly from the mean GC content for the entire genome (40.6%). By contrast, in *Drosophila* and in human, the GC contents of the long homogeneous segments are lower than those of their respective genomes.

Previous work showed that genes in *A. mellifera* preferentially occur in low %GC regions of the genome (*S67*). In this respect, *A. mellifera* was unique among all species we have studied thus far (including human, fly, worm, mosquito, yeast, body louse and sea urchin). We were therefore very interested in whether *N. vitripennis*, being a fairly closely-related hymenopteran, shared this trait. In the more GC-rich half of the *N. vitripennis* genome, the GC content domains in which genes occur are slightly lower in %GC than the GC content domains as a whole (Fig. S19). For example, 69% of the genome is contained in GC content domains whose %GC is less than 45, but this fraction of the genome contains 77% of all *N. vitripennis* genes. So, *N. vitripennis* is like *A. mellifera* in its tendency for its genes to occur preferentially in regions lower than the average genome GC content. However, in *A. mellifera* this effect is far more pronounced.

4. Gene Structure Statistics

Gene structures were measured for EST-validated gene models of *Nasonia* (Table S23) and compared to related insects, and two other animals, given in Table S47. Statistics are based on exon locations that tabulate exon, intron and coding exon locations per gene. Nasonia gene structure is found to be similar to *Apis*, and differs from the Dipteran genes, which have lost introns relative to other animals. The differences between these two hymenoptera are partly attributable to use of EST-validated genes for *Nasonia* but not *Apis*. Measures include Genome size in megabases, where the value in parentheses is gene-containing sequence excluding heterochromatin and scaffolds without genes. No. of genes is the count of gene models from the set examined. Gene density is calculated as the sum of coding exon bases / total gene-containing genome bases. Gene length is the span including introns and UTR. CDS size is the coding sequence length without introns or UTRs. Exons/gene and Exon size are count and size of coding exons. Sizes are given as mean in base pairs except for Intron size. Intergenic size is measured from distance between adjacent genes. These statistics have a standard deviation close to the mean, but Intergenic size has a much larger variance.

VI. Comparative Genomics

1. Arthropod comparative gene analysis

We compared the *Nasonia* Gnomon protein coding gene models of *Apis mellifera*, *Tribolium castaneum*, *Drosophila melanogaster*, *Pediculus humanus*, *Daphnia pulex*, and *Homo sapiens*, to identify homologous and orthologous gene relations (Table S48). A summary of gene orthology and parology for *Nasonia*

and related insects is provided in Table S49. Overall, *Nasonia* encodes a typical insect gene repertoire; 11,579 genes have recognizable orthology, of which almost 6935 genes have a human ortholog. As many as 637 RefSeq genes show no significant homology to the other species, and 927 RefSeq genes could not confidently be assigned orthology despite the presence of conserved domains. *Nasonia* shares with *Apis* 65% median amino acid identity between 1:1 orthologs, significantly higher than to any other species, and comparable to *Aedes-Anopheles* mosquitoes (70%) and chicken-human (73%) comparisons (Fig. S20). Analysis of gene order conservation reveals that 63% of single-copy orthologs are found in micro-synteny blocks for *Nasonia* and *Apis*, versus 75% for *Aedes* and *Anopheles* and 85% for human and chicken (*S154*).

Around 1,000 orthologs appear to have been lost in the *Nasonia* lineage. These include 71 bilaterian specific single-copy orthologs without significant tBlastN match in *Nasonia*. There are 312 orthologs found exclusively in *Nasonia* and *Apis* compared to the other sequenced genomes (Table S50, available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html). Some of these genes likely underpin hymenopteran specific traits such as morphological specializations (the stinger, venom glands, etc), ameiotic spermatogenesis, or parthenogenetic male development. The availability of systemic RNAi opens up an avenue for exploration of their function.

The genome of *Drosophila melanogaster* is known to be considerably derived, with a large number of genes lost (or diverged beyond recognition) that are present in other insects and vertebrates (*S155*). 445 orthologous groups of genes are shared between *Nasonia* and humans, but without a candidate *D. melanogaster* ortholog (Table S1, available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html). Many of these are also found in one or more other invertebrate genomes, including *Tribolium* (323), and *Apis* (380). Ten orthologs are absent in the other invertebrate genomes, but shared exclusively between *Nasonia* and human relative to other invertebrates (Table S39). Examples include a *Nasonia* ortholog of human transcription factors E2F7 and E2F8 (involved in cell-cycle regulation), a potassium channel KCNK13 and serine/threonine kinase 40 (STK40). The discovery in *Nasonia* of genes absent in *Drosophila* opens new avenues for their functional investigation in this genetically tractable insect.

Nasonia has a number of unusual “runaway” amplifications of gene families (Table S40). These include odorant and chemosensory receptor genes, male sex genes, glucosyl transferases, dehydrogenases, and a venom gene family. Table S51 lists a subset of the gene families with orthology where *Nasonia* appears to have overabundant copies, compared to other insects. There are six diverged *Nasonia* genes orthologous to Histone-lysineN-methyltransferase Suv4-20, found as a single copy in most organisms and involved in heterochromatin-induced gene silencing. Another example is an F-box-like/WD protein with four copies in *Nasonia* and only a single human ortholog TBL1K, which plays a crucial role in transcription mediated by nuclear receptors. The *Nasonia* genome encodes a similar but independent expansion of Royal-Jelly-like proteins (Fig.

S5). *Nasonia* also has over 205 proteins bearing Ankyrin (ANK) repeats, a protein motif important for protein-protein interactions (*S156*). This is by far the highest count of ANK proteins in any sequenced arthropod (Table S3). *Nasonia* has 62 genes that code for cuticular proteins with a known chitin binding domain, only about half of the number present in *Anopheles*, *Bombyx* and *Drosophila*, but about twice those found in *Apis*, perhaps reflecting the protected larval environments in which these Hymenoptera develop (Table S11).

2. KEGG analysis

Mapping these gene losses onto metabolic pathways (*S157, S158*) indicates rearrangement or disablement of amino acid metabolic pathways, including tryptophan and aminosugar metabolism, and nuclear-encoded urea cycle components (Figs. S21, S22, S9). Changes to pathways involved in amino acid metabolism may reflect the amino-acid rich carnivorous diet of these parasitoids. Both *Nasonia* and *Apis* are missing key enzymes of aminosugar and tryptophan metabolism, indicating that those losses are ancient. Bees also feed on a protein rich diet (pollen) and the common ancestor of *Nasonia* and *Apis* was likely an insect predator, supporting a possible role of diet in these losses. Knowledge of these gene losses could inform efforts to produce artificial diets for parasitoid mass rearing in biological control and improve hymenopteran cell culture.

3. Evolutionary rates between *Nasonia* and *Apis*

A maximum likelihood framework was used to assess rates of protein evolution between the lineages leading to *Nasonia* and *Apis* and to their common ancestor (either body louse, fly, or beetle). Nuclear encoded mitochondrial (OXPHOS and large ribosomal subunit), immunity (including but not limited to innate immunity and humoral immunity signaling) and circadian proteins are among the fastest evolving in the lineage leading to *Nasonia*, whereas polar granule, innate immunity, humoral immunity signaling, and RNAi proteins evolved fastest in the lineage leading to *Api*s. Relative to the common ancestor, elevated evolutionary rates occur in both *Nasonia* and *Apis* in some mitochondrial processes. Sex determination genes evolve more slowly in *Nasonia*, in line with the hypothesis of a more derived sex determination mechanism in Apis (see below). Wing disc anterior-posterior patterning genes are faster evolving in *Nasonia*, consistent with the highly derived and reduced wing morphology of these parasitoids.

4. Ankyrin repeats and phylogenetic analysis of ANK-PRANC proteins

Using the Pfam-A model for the ankyrin domain (PF00023), we identified 207 proteins bearing an Ank-domain (further referred to as Ank proteins) in *N. vitripennis*. This is the highest count of all sequenced arthropod genomes (Table S3). Ank proteins are involved in protein-protein interaction affecting diverse cellular processes (*S156*), and are of particular interest for the analysis of host-parasite interaction (*S159, S160*). Moreover, previous studies have indicated a connection between the number of Ank proteins and the parasitic lifestyle of the

arthropod endosymbiont *Wolbachia* (*S159*). In particular, CI (cytoplasmic incompatibility) seems to be the result of changes to the cell cycle regulation (*S161*), and hence might be influenced by proteins containing Ank repeats.

A closer analysis of the Ank proteins of *N. vitripennis* shows that 21 of the 207 proteins (OGS v1.0) share more similarity to each other than to any other arthropod Ank protein, suggesting that they have propagated by repeated gene-duplication events. 13 of these 21 Ank proteins have a C-terminal region which matches the Pfam definition of the PRANC domain (PF09372) domain. The PRANC (Pox proteins repeats of ANkyrin - C terminal) domain may be a viral counterpart to the eukaryotic F-box domain (*S162*) as it mediates interaction to ubiquitin-kinases, thereby allowing pox-viruses to exploit the proteasome machinery of their hosts (*S163*).

In order to assure that the PRANC sequences were actually present in the *Nasonia* genome and not viral or bacterial contaminant sequence, we identified mate pairs in which one read matched a PRANC gene and one read matched a non-PRANC gene. We mapped 393 *N. vitripennis* trace sequences to PRANC genes, and 129 of these reads had mates which hit non-PRANC genes. Twelve non-PRANC genes were hit by 5 or more mates, giving high confidence that the PRANC genes are physically linked to other *Nasonia* genes. All 12 of the genes had the highest homology to an insect gene when BLASTed against the NCBI non-redundant protein database, which strongly supports the hypothesis that PRANC genes are *Nasonia* genes and not viral or bacterial contaminant sequence.

As the name suggests, PRANC occurs at the C-terminus of Ankyrin(ANK)-repeat proteins. While PRANC has so far only been identified and described in pox-viruses of the family *Chordopoxvirinae* (vertebrate pox-viruses), we identified thirteen ANK-PRANC genes in the genome of *Nasonia vitripennis.*

We used HMM-based methods to identify possible remote homologs to the PRANC domain in all proteins stored in the GenBank protein non-redundant database, where we found, in total, more than 300 significant hits. While most of these proteins expectedly belong to pox-viruses (92%), we also identified one ANK-PRANC protein in the parasitoid wasp *Cotesia congregata,* and multiple ANK-PRANC proteins in bacteria of the genus *Rickettsia*. Specifically, PRANC has been found in the *Wolbachia* symbionts of *D. simulans* (wSim, two hits) and *D. ananassae* (wAna, 2 hits), in *Orientia tsutsugamushi* (strains Ikeda (13 hits) and Boryong (5 hits)) and in *Rickettsiella grylli* (1 hit).

Using primers based on a PRANC gene from the *Wolbachia* symbiont of *D. ananassae*, we amplified and sequenced ANK-PRANC genes from both A and B group *Wolbachia* (*w*VitA and *w*VitB) from *Nasonia vitripennis* and from a divergent F-group *Wolbachia*.

The presence of multiple copies of ANK-PRANC proteins in *Rickettsia* is unusual, as Ankyrin domains are rare in prokaryotes (*S164*). However, previous studies have shown unusually high numbers of Ankyrin proteins in *Rickettsia*. Furthermore, 60 Ankyrin domain encoding genes (*S164*) have been shown to be present in the genome of the *Wolbachia* endosymbiont of *Culex pipiens* and 50 Ankyrin encoding genes have been reported for *Orientia tsutsugamushi* strain

Boryong (*S165*). The abundance of Ankyrin genes in symbiotic bacteria is one reason Ankyrin genes are seen as playing an important role in host-parasite interactions (*S159*)

PRANC domain sequences in *Rickettsia* and *Nasonia* show substantial divergence from pox-viral PRANC domains, and this might explain why the presence of PRANC has so far not been reported for the above mentioned genomes. The PRANC domain from *Cotesia congregata* is the only known eukaryotic instance of PRANC outside the *Nasonia* genome.

A phylogenetic analysis of PRANC domain sequences was run with a representative subset, using only the domain sequence for a phylogenetic analysis as the N-terminal region of ANK-PRANC proteins is very divergent in both copy number and sequence of the Ankyrin repeats. The PRANC sequences have been aligned and the alignment was used for a Maximum Likelihood analysis (see phylogenetic tree in Fig. 4).

With rare exceptions, the PRANC sequences fall into 4 well-separated clades corresponding to sequences from pox-viruses, *Orientia tsutsugamushi*, *Wolbachia* and *Nasonia vitripennis*. The *Wolbachia* PRANC sequences and the *Cotesia congregata* PRANC sequence appear to relate more closely to *Nasonia* than to pox-viral sequences. There are three pairs of PRANC sequences with 100% sequence identity. The *w*VitA and *w*Vit B PRANC sequences are identical to each other, and each of the two PRANC sequences encoded by wAna has an identical counterpart in *w*Ri. It is significant that one of the *w*Ana/*w*Ri PRANC sequences is more similar to the *Nasonia* PRANC sequences, while the other is more similar to the rest of the *Wolbachia* PRANC sequences.

Most of the *Nasonia* genes with PRANC domains are expressed in all stages of development tested, strongly suggesting function (Table S15).

Our analysis suggests that an ancestor of *Nasonia vitripennis* acquired a PRANC containing gene by a lateral gene transfer, possibly involving *Wolbachia.*

VII. Evolutionary Genetics and Speciation

1. Evolutionary Rates between Nasonia Species

In order to examine evolutionary rates between *Nasonia* species pairs, we estimated dN/dS for all RefSeq genes and then identified GO (Gene Ontology) terms with significantly elevated dN/dS relative to a random distribution (see methods). All GO terms with significantly elevated dN/dS at the 0.01 level for at least one species pair are shown below in Table S13. Genes involving the structure of the mitochondrial ribosome and ATP synthase complex have significantly elevated rates for all species pairs, while genes involving the mitochondrial respiratory chain I (NADH dehydrogenases) are rapidly evolving only between *N. vitripennis* and the other two species.

We further examined dN/dS values for a variety of gene sets using Mann-Whitney U tests, and the results are shown in Table S52. We compared dN/dS ratios between *N. vitripennis* and *N. giraulti* venom genes to non-venom genes (n=43 venom genes, 7,248 non-venom genes, U=226,867). Venom genes

showed significantly higher dN/dS ratios between *N. vitripennis* and *N. giraulti* than non-venom genes (p<2x10-6). We also compared dN/dS ratios between *N. vitripennis* and *N. giraulti* heterochromatic genes to euchromatic genes (n=853 heterochromatic genes, 6,432 euchromatic genes, U=3,066,015). Heterochromatic genes showed a large increase in average dN (0.0081 vs 0.0067) and a small increase in average dS (0.0326 vs 0.0317) relative to euchromatic genes. Genes located in heterochromatin had significantly higher dN/dS ratios between *N. vitripennis* and *N. giraulti* than genes located in euchromatin ($p<2x10^{-5}$). Heterochromatic genes showed a 21% increase in average dN but only a 3% increase in average dS relative to euchromatic genes. Finally, we compared dN/dS ratios between *N. vitripennis* and *N. giraulti* for Hymenoptera-specific genes and *Nasonia*-specific genes to non-specific genes (Hymenoptera-specific: n=147 hymenoptera-specific, n=6774 non- specific, U=648,014; *Nasonia*-specific: n=434 *Nasonia*-specific, U=1,794,197). Hymenoptera-specific and *Nasonia*-specific genes had significantly higher dN/dS ratios than non-specific genes (Mann-Whitney U test, $p<2x10^{-6}$ for Hymenoptera-specific and $p<7x10^{-6}$ for *Nasonia*-specific genes). We also compared *Nasonia*-specific genes vs Hymenoptera-specific genes (U=33,049), which did not have significantly different dN/dS ratios from each other (Mann-Whitney U test, p=0.51).

For *Nasonia*-specific, Hymenoptera-specific, and non-specific genes, we examined how dN/dS compared to gene size distribution (Table S53). Hymenoptera-specific and non-specific genes are mostly large (>65% >1000 bp) while *Nasonia*-specific genes are mostly small (80% <1000 bp). *Nasonia*-specific and Hymenoptera-specific genes had a higher proportion of genes with dN/dS > 1 for all size categories, indicating that the significant differences in dN/dS are not due to differences in gene size between categories.

VIII. Biological Processes

1. Lateral Gene Transfers (LGT)

Nasonia species were antibiotically cured from infection of the endosymbiont Wolbachia prior to sequencing. This facilitated detection of lateral gene transfers (LGTs) from these bacteria to the wasp nuclear genome. Each of the sequenced *Nasonia* species was shown to have recently acquired small fragments of *Wolbachia* genes (*S166*). The number of inserts identified – 4 in *N. vitripennis*, 1 in *N. giraulti*, and 1 in *N. longicornis* – is correlated to the sequencing coverage. This suggests that additional inserts may remain unidentified within the genomes of *N. giraulti* and *N. longicornis*, and in the unsequenced *N. oneida*. Further, sequencing of multiple strains showed that each Wolbachia LGT is species-specific and fixed for each species (*S166*). This fact was corroborated for *N. oneida*, which lacks the insert seen in its sister species, *N. giraulti*.

3. Immunity

Generally, the *Nasonia* immune repertoire is similar to that of *Apis.* The overall gene count in Nasonia is 10-20% higher, while still one-half that known from flies. The inferred *Nasonia* Peptidoglycan Recognition Protein LC (PGRP-LC) locus is the most complex so far found among sequenced insects. It encodes ten PGRP domains that are orthologous to a single domain in *Apis* PGRP-LC, and shows evidence of extensive splice variation (Fig. S7). Exons upstream from the N-terminal PGRP domain encode a PGRP-LC-like cytoplasmic motif and a transmembrane domain that could potentially be spliced in frame with any of the 10 PGRP domains. However, similar to *Drosophila* PGRP-LF, domains 5 and 7-10 have short upstream exons that encode signal peptides or N-terminal membrane anchors, with support in EST and tiling array data.

Among other immune pathway signaling molecules, three cactus homologs are syntenic with one of the dorsal genes. This gives some credibility to the hypothesis that the ancestral Rel gene was a composite one, like Relish, which later split into two to generate dorsal-like proteins (*S167*).

4. Immunity and *Wolbachia* interaction

Young adult males and females of a *N. giraulti* strain carrying unusually high bacterial loads for this species (~10 *Wolbachia* copies per host gene copy) and uninfected controls were profiled for expression differences in 11 immunity genes (see Materials & Methods). Results confirmed expression of the annotated genes, and variable expression of *PGRP-LC* domain 7 in response to *Wolbachia* infection (Table S56). Consistent with the annotation and quantitative expression data, *Nasonia* ESTs show that domain 7 is the most highly expressed exon in the locus. We hypothesize that *Wolbachia*-induced regulation of domain 7 is a putative adaptation to interact with ancient and conserved features of the insect innate immune response.

5. Xenobiotics

Nasonia does not have the elaborate social chemical communication of bees, but might need to detect a diversity of chemicals to find potential hosts or to avoid harmful substances in its environment. *Nasonia* frequents habitats (decomposing carcasses and bird nests) with a variety of noxious compounds. Its genome is well endowed with genes involved in detoxification (see Table S54 and Fig. S8). For example, there are twenty glutathione-S-transferase (GST), including an unusually high number (8) from the Sigma class, which protects against reactive oxygen species, and possibly reflects a parasitoid adaptation for surviving the chemically hostile environment. *Nasonia* has 92 cytochrome P450 genes, with members of the CYP3 and CYP4 classes being particularly abundant compared to *Apis* (49 vs 28 and 29 vs 3, respectively). Twelve of 41 esterases in *Nasonia* are located in two hymenopteran-specific clades, one of which also contains an esterase associated with insecticide resistance in another wasp (*S168*). Overall the GST, P450 and esterase gene complements of *N. vitripennis*

are much more similar to those in sequenced genomes from other insect orders than they are to those in the honeybee.

6. Pathogens, Symbionts, and Immunity

a. Immune-related Gene Repertoire

Total numbers of immune-related genes identified by Hidden Markov Model (=HMM) varied among the five different *Nasonia* gene prediction sets: RefSeq (149), GLEAN (157), Gnomon (167), Fgenesh (172), and Augustus (176). These figures indicate a slightly higher count (13-40 more genes) than the total number of genes identified in *Apis* (136) and a significantly lower count (about half) than the totals found in *Anopheles* (317) or *Aedes* (346). The cross-referencing procedure with the 'known' immune-repertoires of the two mosquitoes and the honeybee indicate that in general, given good gene models, the HMM library performs very well, with the diversity of the automatically defined set complementing the Dipteran bias in the manually curated set. The HMMs perform poorly on the groups antimicrobial peptides, mainly because these are often species- or lineage- specific or they have not been predicted in any of the five *Nasonia* sets (automated prediction of such genes is particularly challenging), or their short sequence lengths mean the scans fail to identify significant matches. Matches to antimicrobial peptides were limited to the relatively long and conserved defensins and peptides with plausible orthologs in *Apis.* The advantage of analyzing several different gene prediction sets is clear as key immune components that appear to be missing in some cases may be found in the other gene sets. Alternatively, manual checking identified likely genomic locations of such genes that were not predicted in any of the five different gene sets (mainly due to sequence gaps in these regions precluding determination of the full gene structure). In general, the *Nasonia* immune repertoire closely mirrors that of the honeybee, where members of most immune gene families are present, including orthologs of the majority of pathway components, while the total repertoire size is much smaller than in the Dipterans.

b. Symbionts

Nasonia has a diverse microbiota that could impact evolution of its immunity. Over 6 different *Wolbachia* species (*S106*) and the male-killing bacterium *Arsenophonus nasoniae* (*S169*) are present in *Nasonia*. *Wolbachia* are intracellular bacteria that have been implicated in reproductive isolation among *Nasonia* species (*S3*). The gamma proteobacterial genus *Arsenophonus* was first described in *Nasonia*, but is now known to be one of the most widespread bacteria in insects (*S170*). *Arsenophonus nasoniae* carries homologues of type III secretion system effectors that are likely to interfere with innate immune function (*S171*)..

The *Nasonia* genome also revealed an apparent commensal bacterium. A number of bacterial sequences assembled in 319 scaffolds made up of a small number of reads and with low coverage (Table S55, available from supporting

data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html). Most of the contaminant bacterial sequences belong to a single Gammaproteobacteria of the genus *Proteus*, and a likely commensal of *Nasonia*. They show 96% average DNA identity with the reference *Proteus mirabilis* genome (*S172*). *Photorhabdus luminescens* normally gives the second best match. These 319 scaffolds include approximately 500 ORFs or partial ORFs and sum up over 400 kb, what is about one-tenth of the expected genome size (*Proteus mirabilis* genome = 4.063 Mb). Bacteria of the genus *Proteus* are known to be associated and probable symbionts with fleshflies and blowflies, common hosts of *Nasonia* species (*S173*). *Nasonia* may well get exposed to it and thus have it as a commensal.

7. Developmental Genetics

The *Nasonia* embryo presents an excellent system to study the evolution of developmental mechanisms as it shares a long germ-band mode of development with *Drosophila*, albeit with an independent evolutionary origin (*S174*). Genes involved in antero-posterior patterning are quite similar in *Nasonia* and *Drosophila* (Fig. S4F, G), although *Nasonia* appears to rely more extensively on localized maternal mRNA for the earliest step of development (*S175*-*S177*) (Fig. S4, A-E). To accommodate their common long germ mode of development, and thus the requirement for extended patterning information, *Drosophila* and *Nasonia* utilize morphogenetic centers at both poles (*S174, S178*) *Nasonia* lacks the *bicoid* gene, the anterior determinant that patterns head and thorax of the fly embryo. Instead, it uses a cohort of localized maternal factors: *orthodenticle* is anteriorly localized to pattern head and thorax, but it is also localized posteriorly (Fig. S4A, E) where it patterns posterior segments (*S175*). *giant* mRNA is localized at the anterior (Fig. S4B; (*S176*)) where it represses the trunk gene *Krüppel*, while *caudal* mRNA is localized posteriorly to spatially restrict its function (Fig. S4C; (*S177*)). In flies, Bicoid directly represses *Krüppel* and prevents anterior translation of *caudal*. Among the genes required for anterior *bicoid* localization, only *staufen* is present in *Nasonia*, while *exuperantia* and *swallow* are not found. The genes involved in posterior *nanos* localization (Fig. S4D) are conserved in *Nasonia*, including *oskar*, which is not found in *Apis*, *Tribolium*, and *Bombyx*. Genes involved in antero-posterior patterning, gap and pair rule genes (Fig. S4F, G) and downstream members of the segmentation cascade are quite similar in *Nasonia* and *Drosophila*.

In *Drosophila*, *tailless* and *huckebein* control terminal development in response to activation of the Torso pathway. Although *Nasonia* has both *Torso* and *torso-like,* it lacks the ligand, *trunk*, and *tailless* is instead regulated by Orthodenticle (*S179*). As *Apis* lacks both *Torso* and *trunk,* terminal system function appears to have been lost from Hymenoptera entirely.

A mutagenesis screen (*S180*) identified not only mutants that resemble fly segmentation phenotypes, but also phenotypes that have no *Drosophila* counterparts, underscoring the potential in *Nasonia* to discover novel patterning strategies. All major components of the Toll pathway, which determines dorso-ventral axis, are present in *Nasonia*, although there are numerous gene

duplications that are *Nasonia*-specific (e.g. four paralogs of the *Toll* receptor, two of its ligand *spätzle* and two of the downstream *pelle* kinase). BMP signaling is critical for establishing dorso-ventral polarity throughout Metazoa**,** and most components are found in *Nasonia*. Orthologs of vertebrate genes absent from *Drosophila* are also found, including the TGFbeta ligands *ADMP* and *myostatin*, and the BMP inhibitors *BAMBI* and *DAN*. The BMP pathway seems to be a hotbed of *Nasonia*-specific gene duplication, including two paralogs each of the ligand *glass bottom boat*, the receptor *punt*, and the co-SMAD *medea* and three paralogs of the BMP modulator *crossveinless*, and four of the inhibitor *crossveinless-2*. Finally, the dorso-ventral fatemap, as illustrated by genes such as *twist* and *vnd* is largely conserved in *Nasonia* (Fig. S4, H).

8. Proteomic Characteristics of diapause in *N. vitripennis*

Following a re-analysis of the mass spectrometric data from that experiment using the newest release of the official protein set (OGS v1.2), we were able to match peptides to 640 of the 18,822 predicted proteins (Table S12, available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html).

9. Sex Determination

Hymenoptera have different mechanisms of haplodiploid sex determination (*S181*), and how this is achieved has puzzled biologists for many decades. In *Apis*, sex is determined by the allelic state of a single locus, the *complementary sex determiner* (*csd*) (*S182*) However, *Nasonia* does not have *csd* (*S183*). Functional studies support a specific role in sex determination for both *doublesex* (*dsx*) (*S184*) and *transformer* (*tra*) (see below). The *Nasonia* primary sex determining system remains to be resolved, but genetic crosses and *tra* knockdown experiments indicate a combination of maternal effects and parental imprinting.

Short conserved repeats, similar to those on *Drosophila doublesex* and *fruitless* primary transcripts and bound by tra/tra2 to achieve female-specific splicing (*S185, S186*), are found in both *N. vitripennis* and *A. mellifera doublesex* and *fruitless* female-specific exons (*S187*), suggesting a similar splicing mechanism in both hymenopteran species (for an overview of sex determination see Fig. S6).

Genomic structure of Nvtra: 5'- and 3'-RACE-PCR of male and female mRNA followed by sequencing yielded the one female and three male specific complete transcripts. As shown in Fig. S6, the complete genomic region of transformer is 7202 bp. The gene is composed of nine exons of which exon 2 is differentially (cryptically) spliced in females and males. The transcription start site is -300 bp from the ATG-start codon located in exon 1. In females a single transcript is produced from exon 1 – 9. Differential splicing of exon 2 causes only the first 188 bp to be included in the female-specific transcript, that codes for a full length protein of 405 aa. In males three different transcripts are produced (M1 – M3), that include either the entire exon 2 (M1), the second 232 bp part of exon 2 (M2) or the last 54 bp (M3) of exon 2 (Fig. S6). All male transcripts yield truncated and

probably non-functional proteins of either 176 or 182 aa due to the presence of one (M3), five (M2) or six (M1) in frame stop codons in the transcript. This indicates that this gene is spliced in a similar way as described for other *tra* genes and the *fem* gene of the honeybee.

Synteny analysis of Apis fem with Nasonia tra: The protein sequences from the genes within the sex determination locus (SDL) from *Apis mellifera* (*S188*) were used in a tBLASTn search to identify orthologs in the Nasonia genome using the HGSC BLAST service with *Nasonia vitripennis* version 1.0. *Nasonia* orthologs of *Apis* genes GB11211 (2.05989 e^{-91}) and GB13727 (5.17382 e^{-38}) were all found within scaffold 116. However, *Nasonia tra,* (*Nvtra)* which is an ortholog of *Apis fem,* is not located in this scaffold. *Apis* gene GB30480 that is also located in the *Apis* SDL, has a *Nasonia* ortholog (4.24421 e^{-07}) located at approximately bp 7000 of scaffold 8. Scaffold mapping (*S38*) indicates that scaffold 116 and 8 are adjacent on chromosome 2. Both *Apis csd* (0.000193447) and *Apis fem* (1.80267 e^{-07}) show homology to a region at bp 700,000 in scaffold 8. This is the location of *Nvtra.* Homologs of two putative genes located within a 58 kb region of the *Apis* genome were identified in scaffold 8, upstream of *Nvtra* in the *Nasonia* genome using tBLASTn. (GB13000 2.02959 e^{-33} and GB16089 1.19287 e^{-98}). Two other putative genes from the same 58 kb *Apis* genomic region were identified in scaffold 8, downstream of *Nvtra* (GB16151 5.93434 e^{-21} and GB13465 1.55567 e^{-148}). These results corroborate the suggestion by Hasselmann et al. (*S188*) that *csd* is not the ancestral hymenopteran sex determining gene in Hymenoptera. They also underline the dynamic nature of sex-determining genes, that not only show duplication but also translocation events (Fig. S23)

10. Sex ratio control

Nasonia is a classic model for adaptive sex ratio evolution (*S189*): The haplodiploid mode of sex determination allows parasitoid wasps to control the sex ratio among progeny, because females can determine whether individual eggs are fertilized (resulting in daughters) or not (resulting in sons). Utilizing markers derived directly from the genome sequence, candidate sex ratio control QTL have been identified (see below), and a locus that disrupts sex ratio control in hybrids has been mapped (see below). Identifying genes behind sex ratio control is relevant both to evolutionary theory and for improvement of mass-rearing of parasitoids for biological pest control.

The chromosomal region encompassing the *st-DR* locus was introgressed from *N. giraulti* into *N. vitripennis* by backcrossing wild-type eye colored females to *st-DR* (red eyes) males for a total of 15 generations before being made homozygous. The resulting line, Int-$Stdr_g$, exhibits an aberrant sex ratio, manifest only in females. The average proportion female was 0.30 (±0.03 SE, N=43) and 0.43 (± 0.04, N=45) when Int-$Stdr_g$ females were mated to Int-$Stdr_g$ or AsymCX males compared to 0.84 (± 0.02, N=31) when AsymCX females were mated to Int-$Stdr_g$ males or 0.88 (± 0.01, N=27) for AsymCX females mated to AsymCX males. Reduced average family size seen in the introgression line (49.51 ± 3.57 for Int-$Stdr_g$ x Int-$Stdr_g$ versus 62.89 ± 4.71 for AsymCX x AsymCX) is not

sufficient to alone explain the reductions in females. The *N. giraulti* chromosomal segment introgressed into *N. vitripennis* results in a 74% reduction in average number of female offspring compared to only 21% reduction in family size, indicating a hybrid disruption in sex ratio control.

A quantitative trait loci (QTL) study on naturally-occurring intra-specific variation in sex ratio in *Nasonia vitripennis* (*S189*) identified two putative QTL, one on chromosome two and one on chromosome three. In addition, this study also identified a putative QTL for brood size on chromosome one.

11. Neurohormones and ion channels

Biogenic amines, neuropeptides and protein hormones (collectively called neurohormones) and their G-protein coupled receptors (GPCRs) direct central processes in insects such as development, reproduction, and behavior (*S190*). We found 15 biogenic amine GPCRs in *Nasonia* and 39 neuropeptide and protein hormone GPCRs (Table S9). In addition, we identified 29 neuropeptide and protein hormone genes in *Nasonia* (Table S9). *Nasonia* has the largest number of cys-loop ligand-gated ion channels found in insects (26 genes), which mediate the fast neurotransmission (*S191*), and numerous DEG/ENaC and TRP ion channels, which are involved in sensory information processing.

Nasonia, together with *Tribolium castaneum*, is the second insect species known to have a gene coding for a hormone homologous to vertebrates oxytocin and vasopressin. In vertebrates, this gene has been duplicated from an ancestral ortholog during the transition from fishes to tetrapods (*S192*). A phylogenetic reconstruction was made to understand the relationship of the insect ortholog – called inotocin (*S193*) – to other oxytocin-/vasopressin-like genes (Fig. S24). Duplicated versions of the ancestral gene are apparent in vertebrates, where the most studied counterparts are oxytocin and vasopressin (arginine vasopressin, AVP) in mammals. In vertebrates, class-specific clusters (e.g. birds, mammals) are generally strongly supported. Furthermore, vertebrates are clearly separated from invertebrates. The inotocin genes from *N. vitripennis* and *T. castaneum* form a distinct cluster (probably insect-specific) clearly separated from vertebrate but also from other invertebrate orthologs. In *Nasonia*, inotocin is expressed throughout development and shows testes-specific splice variation, but its function remains to be established.

DEG/ENaC and TRP channels in *Drosophila* are involved in the perception of sensory signals such as light, temperature, humidity, pheromones, sound, and touch (*S194, S195*). The *Drosophila* genome contains at least 20 DEG/ENaC genes, whereas only 5 such genes could be detected in *Nasonia* (Table S10). It is interesting that the honey bee also contains 5 DEG/ENaC genes that appear to be the direct orthologs of the *Nasonia* genes. In addition, *Nasonia* contains 12 TRP channel genes (13 in *Drosophila*; Table S10). The honey bee contains 11 TRP channel genes that seem to be the orthologs of the Nasonia genes. Thus, the number of DEG/ENaC and Trp channels seems to be well conserved in Hymenoptera, whereas the number of DEG/ENaC channels has strongly expanded in Diptera.

12. Courtship

The genetic bases of speciation and species differences and whether these differences have evolved by natural selection or drift are major themes in evolutionary biology. In this study we analyzed the genetic architecture of species differences in male courtship behavior between *N. giraulti* and *N. longicornis,* which can still interbreed once they are cured from their endosymbiont *Wolbachia*. This analysis provided us with valuable information both about the genetic architecture (number and effect of loci involved in species differences) and the process how these differences came about (drift versus selection).

We found eight QTL, that were significant on a genome wide level of p=0.01 (Fig. S25 and Table S57). Six of these eight QTL were associated with “number of headnods” in cycle 1-4 and the same marker on chromosome 1 (Scaf127_676512) was significant for “number of headnods in cycle 1, 2 and 4” (= three QTL) and marker Scaf176_257755 on chromosome 5 accounted for two other QTL (Fig. S25 and Table S57). Marker Scaf1_3410194 on chromosome 5 was associated with the last of the six QTL for “number of headnods-3rd cycle” and could reflect an independent 3 QTL for “number of headnods” (Table S57). It is likely that the three QTL on chromosome 1 and the two QTL on chromosome 5 have the same genetic basis since they influence the same trait “head nodding” but at a different time during the male courtship behavior (cycle 1,2+4 and 1+2, respectively) Two QTL were found for number of cycles (Fig. S25 and Table S57). No significant QTL were found for latency and all four measured cycle times.

N. longicornis and *N. giraulti* are allopatric, i.e. any phenotypic and the underlying genotypic differences have most likely evolved by drift and not selection (e.g. reinforcement). We found at least four QTL two for headnods on chromosome 1 and 5 and two for total number of cycles on chromosome 2 and 4 (Fig. S25 and Table S57), that explained a large proportion (>10%) of the observed phenotypic variance in our mapping population. Although both species have separated relatively recently there was enough time to accumulate fixed differences for male courtship components that could contribute to prezygotic isolation if these species would become sympatric again.

References

1. J. A. J. Breeuwer, J. H. Werren, *Evolution* **49**, 705 (1995).
2. J. H. Werren, *Annu. Rev. Entomol.* **42**, 587 (1997).
3. S. R. Bordenstein, F. P. O'Hara, J. H. Werren, *Nature* **409**, 707 (Feb 8, 2001).
4. P. Havlak *et al.*, *Genome Res* **14**, 721 (Apr, 2004).
5. A. Coghlan *et al.*, *BMC Bioinformatics* **9**, 549 (2008).
6. Y. Kapustin, A. Souvorov, T. Tatusova, paper presented at the RECOMB 2004 - Currents in Computational Molecular Biology., 2004.
7. B. Kiryutin, A. Souvorov, paper presented at the ISMB 2005, 2005.
8. A. Souvorov, T. Tatusova, D. Lipman, paper presented at the ISMB 2004, 2004.
9. B. J. Haas *et al.*, *Nucleic Acids Res* **31**, 5654 (Oct 1, 2003).
10. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res.* **35**, D61 (Jan, 2007).
11. D. Maglott, J. Ostell, K. D. Pruitt, T. Tatusova, *Nucleic Acids Res.* **35**, D26 (Jan, 2007).
12. M. Stanke, M. Diekhans, R. Baertsch, D. Haussler, *Bioinformatics* **24**, 637 (Mar 1, 2008).
13. V. Solovyev, P. Kosarev, I. Seledsov, D. Vorobyev, *Genome Biol.* **7 Suppl 1**, S10 1 (2006).
14. C. G. Elsik *et al.*, *Genome Biol* **8**, R13 (2007).
15. A. A. Salamov, V. V. Solovyev, *Genome Res.* **10**, 516 (Apr, 2000).
16. V. Solovyev, in *Handbook of Statistical Genetics,* D. J. Balding, M. Bishop, C. Cannings, Eds. (John Wiley & Sons, Chichester, England; Hoboken, NJ, 2007), pp. 97-159.
17. M. Stanke, S. Waack, *Bioinformatics* **19 Suppl 2**, ii215 (Oct, 2003).
18. A. Bairoch, B. Boeckmann, S. Ferro, E. Gasteiger, *Brief. Bioinform.* **5**, 39 (2004).
19. G. S. Slater, E. Birney, *BMC Bioinformatics* **6**, 31 (2005).
20. T. D. Wu, C. K. Watanabe, *Bioinformatics* **21**, 1859 (May 1, 2005).
21. W. R. Pearson, D. J. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2444 (Apr, 1988).
22. S. E. Lewis *et al.*, *Genome Biol* **3**, RESEARCH0082 (2002).
23. W. B. Hunter *et al.*, *J Insect Sci* **3**, 23 (2003).
24. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (Oct 5, 1990).
25. Z. Tang *et al.*, *BMC Genomics* **10**, 174 (2009).
26. W. J. Kent, *Genome Res* **12**, 656 (Apr, 2002).
27. B. M. Bolstad, R. A. Irizarry, M. Astrand, T. P. Speed, *Bioinformatics* **19**, 185 (Jan 22, 2003).
28. E. V. Kriventseva, N. Rahman, O. Espinosa, E. M. Zdobnov, *Nucleic Acids Res* **36**, D271 (Jan, 2008).

29. S. Griffiths-Jones, H. K. Saini, S. van Dongen, A. J. Enright, *Nucleic Acids Res* **36**, D154 (Jan, 2008).
30. K. Katoh, H. Toh, *BMC Bioinformatics* **9**, 212 (2008).
31. E. Bonnet, J. Wuyts, P. Rouze, Y. Van de Peer, *Bioinformatics* **20**, 2911 (Nov 22, 2004).
32. B. Giardine *et al.*, *Genome Res* **15**, 1451 (Oct, 2005).
33. D. Gerlach, E. V. Kriventseva, N. Rahman, C. E. Vejnar, E. M. Zdobnov, *Nucleic Acids Res* **37**, D111 (Jan, 2009).
34. I. L. Hofacker, *Curr Protoc Bioinformatics* **Chapter 12**, Unit 12 2 (Feb, 2004).
35. C. Notredame, D. G. Higgins, J. Heringa, *J. Mol. Biol.* **302**, 205 (Sep 8, 2000).
36. W. R. Pearson, *Methods Enzymol.* **183**, 63 (1990).
37. Z. Zhang, S. Schwartz, L. Wagner, W. Miller, *J Comput. Biol.* **7**, 203 (Feb-Apr, 2000).
38. O. Niehuis *et al.*, *PLoS ONE* DOI: 10.1371/journal.pone.0008597 (Jan, 2010).
39. B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res* **8**, 175 (Mar, 1998).
40. X. Wang, B. Seed, *Bioinformatics* **19**, 796 (May 1, 2003).
41. D. W. Loehlin, L. S. Enders, J. H. Werren, *Heredity* DOI: 10.1038/hdy.2009.146 (Jan, 2010).
42. R. C. Edgar, E. W. Myers, *Bioinformatics* **21 Suppl 1**, i152 (Jun, 2005).
43. R. C. Edgar, *BMC Bioinformatics* **5**, 113 (Aug 19, 2004).
44. A. F. A. Smit, R. Hubley, P. Green. (1996-2004), pp. RepeatMasker Open-3.0.
45. G. Benson, *Nucleic Acids Res.* **27**, 573 (Jan 15, 1999).
46. R. D. Finn *et al.*, *Nucleic Acids Res* **36**, D281 (Jan, 2008).
47. D. G. Eickbush, T. H. Eickbush, J. H. Werren, *Chromosoma* **101**, 575 (Aug, 1992).
48. J. Jurka *et al.*, *Cytogenet Genome Res* **110**, 462 (2005).
49. C. D. Smith *et al.*, *Gene* **389**, 1 (Mar 1, 2007).
50. M. A. McClure *et al.*, *Genomics* **85**, 512 (Apr, 2005).
51. W. Gish, *http://blast.wustl.edu*, (1996-2004).
52. Z. Xu, H. Wang, *Nucleic acids research* **35**, W265 (Jul 1, 2007).
53. C. D. Smith, S. Shu, C. J. Mungall, G. H. Karpen, *Science* **316**, 1586 (Jun 15, 2007).
54. H. M. Robertson, K. H. Gordon, *Genome Res.* **16**, 1345 (Nov, 2006).
55. M. Osanai, K. K. Kojima, R. Futahashi, S. Yaguchi, H. Fujiwara, *Gene* **376**, 281 (Jul 19, 2006).
56. R. Frydrychova, P. Grossmann, P. Trubac, M. Vitkova, F. Marec, *Genome* **47**, 163 (Feb, 2004).
57. M. Vitkova, J. Kral, W. Traut, J. Zrzavy, F. Marec, *Chromosome Res.* **13**, 145 (2005).
58. H. Fujiwara, M. Osanai, T. Matsumoto, K. K. Kojima, *Chromosome Res* **13**, 455 (2005).

59. S. Richards *et al.*, *Nature* **452**, 949 (Apr 24, 2008).
60. M. G. Goll, T. H. Bestor, *Annu Rev Biochem* **74**, 481 (2005).
61. N. Kunert, J. Marhold, J. Stanke, D. Stach, F. Lyko, *Development* **130**, 5083 (Nov, 2003).
62. J. Marhold *et al.*, *Insect Mol. Biol.* **13**, 117 (Apr, 2004).
63. A. Dong *et al.*, *Nucleic Acids Res* **29**, 439 (Jan 15, 2001).
64. R. Maleszka, *Epigenetics* **3**, 188 (Jul-Aug, 2008).
65. R. Albalat, *Dev. Genes Evol.* **218**, 691 (Dec, 2008).
66. C. P. Ponting, *Nature Rev. Genet.* **9**, 689 (Sep, 2008).
67. H. G. S. Consortium, *Nature* **443**, 931 (Oct 26, 2006).
68. M. J. Sharkey, *Zootaxa* **1668**, 521 (2007).
69. D. Grimaldi, M. S. Engel, *Evolution of the Insects*. (Cambridge University Press, Cambridge, U.K., 2005).
70. D. L. J. Quicke, H. H. Basibuyuk, M. G. Fitton, A. P. Rasnitsyn, *Zoologica Scripta* **28**, 175 (1999).
71. J. B. Whitfield, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7508 (May 28, 2002).
72. J. Savard *et al.*, *Genome Res.* **16**, 1334 (Nov, 2006).
73. V. Krauss *et al.*, *Mol. Biol. Evol.* **25**, 821 (May, 2008).
74. B. M. Wiegmann *et al.*, *BMC Biol* **7**, 34 (2009).
75. J. B. Whitfield, K. M. Kjer, *Annu. Rev. Entomol.* **53**, 449 (2008).
76. J. C. Regier *et al.*, *Syst. Biol.* **57**, 920 (2008).
77. J. C. Regier, J. W. Shultz, R. E. Kambic, *Proc. R. Soc. London Ser. B* **272**, 395 (Feb 22, 2005).
78. S. J. Bourlat, C. Nielsen, A. D. Economou, M. J. Telford, *Mol. Phylogenet. Evol.* **49**, 23 (Oct, 2008).
79. P. Bernaola-Galván, R. Román-Roldán, J. L. Oliver, *Physical Review E* **53**, 5181 (1996).
80. N. Cohen, T. Dagan, L. Stone, D. Graur, *Mol. Biol. Evol.* **22**, 1260 (May, 2005).
81. M. Weber *et al.*, *Nature Genet.* **39**, 457 (Apr, 2007).
82. S. Suzuki *et al.*, *PLoS Genet* **3**, e55 (Apr 13, 2007).
83. S. R. Eddy, *Bioinformatics* **14**, 755 (1998).
84. A. Heger, C. A. Wilton, A. Sivakumar, L. Holm, *Nucleic Acids Res* **33**, D188 (Jan 1, 2005).
85. Y. Wurm *et al.*, *BMC Genomics* **10**, 5 (2009).
86. I. M. Wallace, O. O'Sullivan, D. G. Higgins, C. Notredame, *Nucleic Acids Res.* **34**, 1692 (2006).
87. A. Stamatakis, *Bioinformatics* **22**, 2688 (Nov 1, 2006).
88. F. Abascal, R. Zardoya, D. Posada, *Bioinformatics* **21**, 2104 (May 1, 2005).
89. D. T. Jones, W. R. Taylor, J. M. Thornton, *Comput Appl Biosci* **8**, 275 (Jun, 1992).
90. G. Talavera, J. Castresana, *Syst Biol* **56**, 564 (Aug, 2007).
91. S. Guindon, O. Gascuel, *Syst Biol* **52**, 696 (Oct, 2003).
92. A. Subramanian, H. Kuehn, J. Gould, P. Tamayo, J. P. Mesirov, *Bioinformatics* **23**, 3251 (Dec 1, 2007).

93. R. Durbin, S. Eddy, A. Krogh, G. Mitchison, in *Biological sequence analysis: probablistic models of proteins and nucleic acids* (Cambridge University Press, Cambridge, 1998).
94. K. Rutherford *et al.*, *Bioinformatics* **16**, 944 (Oct, 2000).
95. K. Tamura, J. Dudley, M. Nei, S. Kumar, *Mol Biol Evol* **24**, 1596 (Aug, 2007).
96. F. Ronquist, J. P. Huelsenbeck, *Bioinformatics* **19**, 1572 (Aug 12, 2003).
97. J. C. Wilgenbusch, D. Swofford, *Curr Protoc Bioinformatics* **Chapter 6**, Unit 6 4 (Feb, 2003).
98. D. Posada, K. A. Crandall, *Bioinformatics* **14**, 817 (1998).
99. R. Chenna *et al.*, *Nucleic Acids Res.* **31**, 3497 (Jul 1, 2003).
100. J. Brennecke *et al.*, *Cell* **128**, 1089 (Mar 23, 2007).
101. N. Elango, S. V. Yi, *Mol Biol Evol* **25**, 1602 (Aug, 2008).
102. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389 (Sep 1, 1997).
103. S. Foret, R. Maleszka, *Genome Res* **16**, 1404 (Nov, 2006).
104. Z. Yang, *Mol. Biol. Evol.* **24**, 1586 (Aug, 2007).
105. J. D. Storey, *J. R. Stat. Soc., Ser. B* **64**, 479 (2002).
106. R. Raychoudhury, L. Baldo, D. C. Oliveira, J. H. Werren, *Evolution* **63**, 165 (Jan, 2009).
107. T. A. Hall, *Nucleic Acids Symposium Series* **41**, 95 (1999).
108. J. Rozas, J. C. Sanchez-DelBarrio, X. Messeguer, R. Rozas, *Bioinformatics* **19**, 2496 (Dec 12, 2003).
109. F. Wolschin, G. Gadau, *PLoS ONE* **4: e6394**, (2009).
110. L. Y. Geer *et al.*, *J Proteome Res* **3**, 958 (Sep-Oct, 2004).
111. N. Peiren *et al.*, *FEBS Lett* **580**, 4895 (Sep 4, 2006).
112. F. Vanrobaeys, R. Van Coster, G. Dhondt, B. Devreese, J. Van Beeumen, *J Proteome Res* **4**, 2283 (Nov-Dec, 2005).
113. K. A. Reidegeld *et al.*, *Proteomics* **8**, 1129 (Mar, 2008).
114. R. M. Waterhouse *et al.*, *Science* **316**, 1738 (Jun 22, 2007).
115. A. Bateman *et al.*, *Nucleic Acids Res.* **27**, 260 (Jan 1, 1999).
116. J. D. Evans *et al.*, *Insect Mol Biol* **15**, 645 (Oct, 2006).
117. M. W. Pfaffl, *Nucleic Acids Res* **29**, e45 (May 1, 2001).
118. J. van de Assem, in *Insect Parasitoids, 13th Symposium of the Royal Entomological Society of London,* J. K. Waage, D. J. Greathead, Eds. (Academic Press, London, 1986), pp. 137-167.
119. J. van den Assem, J. H. Werren, *J. Insect Behav.* **7**, 53 (1994).
120. J. W. Van Ooijen, M. P. Boer, C. Jansen, C. Maliepaard. (Plant Research International, Wageningen, the Netherlands, 2002).
121. R. Maleszka, R. Kucharski, *Biochem. Biophys. Res. Commun.* **270**, 773 (Apr 21, 2000).
122. M. D. Drapeau, S. Albert, R. Kucharski, C. Prusko, R. Maleszka, *Genome Res* **16**, 1385 (Nov, 2006).
123. J. E. Rebers, J. H. Willis, *Insect Biochem Mol Biol* **31**, 1083 (Oct, 2001).
124. R. S. Cornman *et al.*, *BMC Genomics* **9**, 22 (2008).
125. R. S. Cornman, J. H. Willis, *Insect Biochem Mol Biol* **38**, 661 (Jun, 2008).
126. M. V. Karouzou *et al.*, *Insect Biochem Mol Biol* **37**, 754 (Aug, 2007).

127. R. Futahashi *et al.*, *Insect Biochem Mol Biol* **38**, 1138 (Dec, 2008).
128. N. He *et al.*, *Insect Biochem Mol Biol* **37**, 135 (Feb, 2007).
129. R. Kucharski, J. Maleszka, R. Maleszka, *Insect Biochem Mol Biol* **37**, 128 (Feb, 2007).
130. T. Togawa, W. Augustine Dunn, A. C. Emmons, J. H. Willis, *Insect Biochem Mol Biol* **37**, 675 (Jul, 2007).
131. X. Guan, B. W. Middlebrooks, S. Alexander, S. A. Wasserman, *Proc Natl Acad Sci U S A* **103**, 16794 (Nov 7, 2006).
132. E. Danty *et al.*, *J Neurosci* **19**, 7468 (Sep 1, 1999).
133. J. Maleszka, S. Foret, R. Saint, R. Maleszka, *Dev Genes Evol* **217**, 189 (Mar, 2007).
134. E. M. Rasch, J. D. Cassidy, R. C. King, *Chromosoma* **59**, 323 (Feb 23, 1977).
135. E. M. Zdobnov, R. Apweiler, *Bioinformatics* **17**, 847 (Sep, 2001).
136. D. E. Stage, T. H. Eickbush, *Insect Mol. Biol.* **19**, 37 (Jan, 2010).
137. T. P. Jurkowski *et al.*, *RNA* **14**, 1663 (Aug, 2008).
138. S. Kumar *et al.*, *Nucleic Acids Res* **22**, 1 (Jan 11, 1994).
139. X. Cheng, R. M. Blumenthal, *Structure* **16**, 341 (Mar, 2008).
140. T. Yokomine, K. Hata, M. Tsudzuki, H. Sasaki, *Cytogenet. Genome Res.* **113**, 75 (2006).
141. N. Elango, B. B. Hunt, M. A. D. Goodisman, S. V. Yi, *Proc Natl Acad Sci U S A* **106**, 11206 (Jul, 2009).
142. M. M. Suzuki, A. R. Kerr, D. De Sousa, A. Bird, *Genome Res* **17**, 625 (May, 2007).
143. G. B. Saul, *Genetics Maps*. S. J. O'Brien, Ed., (Cold Spring Harbor Press, Cold Spring Harbor, NY, 1993).
144. S. L. Ryan, G. B. Saul, 2nd, G. W. Conner, *J Hered* **78**, 273 (Jul-Aug, 1987).
145. F. Perfectti, J. H. Werren, *Evolution* **55**, 1069 (May, 2001).
146. M. J. Perrot-Minnot, J. H. Werren, *Heredity* **87**, 8 (Jul, 2001).
147. A. Bhutkar, S. M. Russo, T. F. Smith, W. M. Gelbart, *Genome Res.* **17**, 1880 (Dec, 2007).
148. M. Ashburner *et al.*, *Nat Genet* **25**, 25 (May, 2000).
149. G. Bernardi, *Gene* **241**, 3 (2000).
150. E. S. Lander *et al.*, *Nature* **409**, 860 (Feb 15, 2001).
151. W. Li, *Gene* **300**, 129 (Oct 30, 2002).
152. W. Li, P. Bernaola-Galvan, P. Carpena, J. L. Oliver, *Comput Biol Chem* **27**, 5 (Feb, 2003).
153. O. Clay, G. Bernardi, *Mol. Biol. Evol.* **22**, 2315 (Dec, 2005).
154. E. M. Zdobnov, P. Bork, *Trends Genet.* **23**, 16 (Jan, 2007).
155. S. Wyder, E. V. Kriventseva, R. Schroder, T. Kadowaki, E. M. Zdobnov, *Genome Biol* **8**, R242 (2007).
156. L. K. Mosavi, T. J. Cammett, D. C. Desrosiers, Z. Y. Peng, *Protein Sci.* **13**, 1435 (Jun, 2004).
157. I. Letunic, T. Yamada, M. Kanehisa, P. Bork, *Trends Biochem. Sci.* **33**, 101 (Mar, 2008).

158. M. Kanehisa *et al.*, *Nucleic Acids Research* **36**, D480 (Jan, 2008).
159. L. R. Serbus, C. Casper-Lindley, F. Landmann, W. Sullivan, *Annu. Rev. Genet.* **42**, 683 (2008).
160. S. R. Bordenstein, M. L. Marshall, A. J. Fry, U. Kim, J. J. Wernegreen, *PLoS Pathog.* **2**, e43 (May, 2006).
161. U. Tram, P. M. Ferree, W. Sullivan, *Microbes Infect.* **5**, 999 (Sep, 2003).
162. S. J. Chang *et al.*, *J. Virol.* **83**, 4140 (May, 2009).
163. S. Sonnberg, B. T. Seet, T. Pawson, S. B. Fleming, A. A. Mercer, *Proc Natl Acad Sci U S A* **105**, 10955 (Aug 5, 2008).
164. T. Walker *et al.*, *BMC Biol* **5**, 39 (2007).
165. N. H. Cho *et al.*, *Proc Natl Acad Sci U S A* **104**, 7981 (May 8, 2007).
166. J. C. Dunning-Hotopp *et al.*, *Science* **317**, 1753 (Sep 21, 2007).
167. M. S. Dushay, B. Asling, D. Hultmark, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10343 (Sep 17, 1996).
168. Y. C. Zhu, A. K. Dowdy, J. E. Baker, *Pesticide Science* **55**, 398 (1999).
169. J. H. Werren, S. W. Skinner, A. M. Huger, *Science* **231**, 990 (Feb 28, 1986).
170. O. Duron *et al.*, *BMC Biol* **6**, 27 (2008).
171. T. Wilkes *et al.*, *Insect Mol. Biol.* **19**, 59 (Jan, 2010).
172. M. M. Pearson *et al.*, *J Bacteriol* **190**, 4027 (Jun, 2008).
173. G. R. Erdmann, *Parasitol Today* **3**, 214 (Jul, 1987).
174. M. I. Rosenberg, J. A. Lynch, C. Desplan, *Biochim. Biophys. Acta*, **1789**, 333-342 (April, 2009).
175. J. A. Lynch, A. E. Brent, D. S. Leaf, M. A. Pultz, C. Desplan, *Nature* **439**, 728 (Feb 9, 2006).
176. A. E. Brent, G. Yucel, S. Small, C. Desplan, *Science* **315**, 1841 (Mar 30, 2007).
177. E. C. Olesnicky *et al.*, *Development* **133**, 3973 (Oct, 2006).
178. P. Z. Liu, T. C. Kaufman, *Evol. Dev.* **7**, 629 (Nov-Dec, 2005).
179. J. A. Lynch, E. C. Olesnicky, C. Desplan, *Dev. Genes Evol.* **216**, 493 (Jul-Aug, 2006).
180. M. A. Pultz *et al.*, *Genetics* **154**, 1213 (Mar, 2000).
181. J. M. Cook, *Heredity* **71**, (1993).
182. M. Beye, *Bioessays* **26**, 1131 (Oct, 2004).
183. S. W. Skinner, J. H. Werren, *Genetics* **94**, 98 (1980).
184. D. C. S. G. Oliveira *et al.*, *Insect Mol. Biol.*, **19**, 99 (Jan, 2010).
185. M. L. Hedley, T. Maniatis, *Cell* **65**, 579 (May 17, 1991).
186. V. Heinrichs, L. C. Ryner, B. S. Baker, *Mol. Cell. Biol.* **18**, 450 (Jan, 1998).
187. R. C. Bertossa, L. van de Zande, L. W. Beukeboom, *Mol Biol Evol* **26**, 1557 (Jul, 2009).
188. M. Hasselmann *et al.*, *Nature* **454**, 519 (Jul 24, 2008).
189. B. A. Pannebakker *et al.*, *Evolution* **62**, 1921 (Aug, 2008).
190. F. Hauser *et al.*, *Front Neuroendocrinol* **29**, 142 (Jan, 2008).
191. A. K. Jones, A. N. Bera, K. Lees, D. B. Sattelle, *Heredity* DOI: 10.1038/hdy.2009.97 (Jan, 2010).

192. R. Acher, J. Chauvet, M. T. Chauvet, *Adv. Exp. Med. Biol.* **395**, 615 (1995).
193. E. Stafflinger *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 3262 (Mar 4, 2008).
194. H. Lin, K. J. Mann, E. Starostina, R. D. Kinser, C. W. Pikielny, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 12831 (Sep 6, 2005).
195. L. Liu *et al.*, *Nature* **450**, 294 (Nov 8, 2007).
196. L. R. Bell, E. M. Maine, P. Schedl, T. W. Cline, *Cell* **55**, 1037 (Dec 23, 1988).
197. J. M. Belote, M. McKeown, R. T. Boggs, R. Ohkawa, B. A. Sosnowski, *Dev Genet* **10**, 143 (1989).
198. K. C. Burtis, B. S. Baker, *Cell* **56**, 997 (Mar 24, 1989).
199. J. W. Erickson, J. J. Quintero, *PLoS Biol* **5**, e332 (Dec, 2007).
200. A. Dubendorfer, M. Hediger, G. Burghardt, D. Bopp, *Int J Dev Biol* **46**, 75 (Jan, 2002).
201. G. Saccone *et al.*, First research coordination meeting, IAEA/FAO, Vienna (1996).
202. A. Pane, M. Salvemini, P. Delli Bovi, C. Polito, G. Saccone, *Development* **129**, 3715 (Aug, 2002).
203. D. Lagos, M. F. Ruiz, L. Sanchez, K. Komitopoulou, *Gene* **348**, 111 (Mar 28, 2005).
204. D. Lagos, M. Koukidou, C. Savakis, K. Komitopoulou, *Insect Mol Biol* **16**, 221 (Apr, 2007).
205. M. F. Ruiz *et al.*, *Genetics* **171**, 849 (Oct, 2005).
206. M. F. Ruiz *et al.*, *PLoS ONE* **2**, e1239 (2007).
207. M. Beye, M. Hasselmann, M. K. Fondrk, R. E. Page, S. W. Omholt, *Cell* **114**, 419 (Aug 22, 2003).
208. S. Cho, Z. Y. Huang, J. Zhang, *Genetics* **177**, 1733 (Nov, 2007).
209. J. D. Bendtsen, H. Nielsen, G. von Heijne, S. Brunak, *J Mol Biol* **340**, 783 (Jul 16, 2004).
210. A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer, *J Mol Biol* **305**, 567 (Jan 19, 2001).

Supplementary Figures

Figure S1: *Nasonia vitripennis* genomic repeat content and gene density distribution.
Figure S2. Distribution of observed to expected CpG ratios in genomes, introns and coding exons for *A. mellifera* (right), *N. vitripennis* (middle) and *D. melanogaster* (right).
Figure S3: Distribution of methylated and unmethylated CpGs in five *N. vitripennis* genes as determined by bisulfite sequencing.
Figure S4: *Nasonia* segmentation genes.
Figure S5: Phylogenetic tree illustrating the relationship between the Yellow/Royal Jelly proteins in microbes, *Drosophila*, *Tribolium*, *Nasonia*, and *Apis*.
Figure S6: The sex determination cascade of Diptera and Hymenoptera.
Figure S7: The major PGRP-LC-like locus in *N. vitripennis* encodes ten PGRP domains.
Figure S8: Detoxification genes mapped onto chromosome 1 and 2 of the *Nasonia* parasitoid wasp.
Figure S9: Kegg diagram showing functional pathways and clusters of gene losses in *Nasonia.*
Figure S10: Consensus qualities of *N. giraulti* and *N. longicornis* aligned sequences.
Figure S11: Read coverage statistics for *N. giraulti* and *N. longicornis* aligned sequences.
Figure S12: Venn diagram depicting 12,811 genes that are validated by expression data from expression tiling arrays and from the large EST sequencing project.
Figure S13: Genomic cluster of microRNA genes conserved between *Nasonia* and honeybee.
Figure S14: The *Nasonia* telomerase or TERT **(A)** gene and **(B)** protein.
Figure S15: Alignment of the catalytic domains of wasp, honeybee and human DNMT1 and DNMT3 orthologs.
Figure S16: Visualization of NvDnmt1a RT-PCR products by ethidium bromide staining after agarose gel electrophoresis.
Figure S17: Genomic organization of selected Yellow/Royal Jelly genes in four insect species: *Drosophila*, *Tribolium*, *Nasonia*, and *Apis*.
Figure S18: Frequency of homogeneous GC domains.
Figure S19: Cumulative distribution of the fraction of nucleotides with GC content in GC content domains containing genes (thick lines) and all GC content domains (thin lines) for *N. vitripennis* and *A. mellifera*.
Figure S20: Distribution of amino acid identity of single-copy orthologs between *Nasonia* and the other species.
Figure S21: *Nasonia vitripennis* aminosugar metabolism.
Figure S22: *Nasonia vitripennis* tryptophan metabolism.
Figure S23: A schematic overview of the synteny between the SDL genomic region in *A. mellifera* and the genomic region of *Nvtra*.

Figure S24: Phylogenetic tree of oxytocin/vasopressin-like genes.
Figure S25. Three QTL for male courtship behavior..

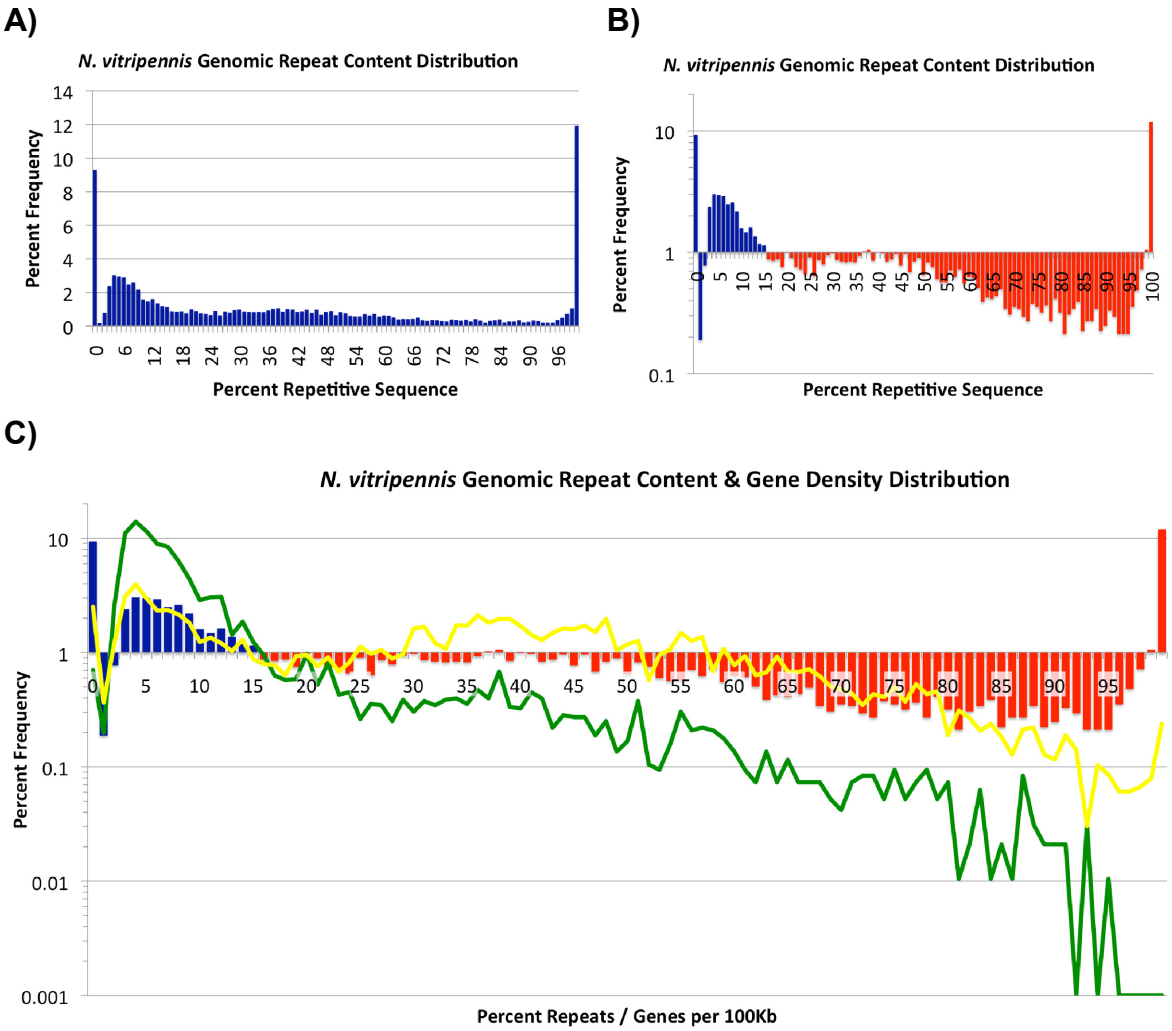


Figure S1: *Nasonia vitripennis* genomic repeat content and gene density distribution. **(A)** Large scaffolds were segmented into 100kb or smaller fragments and the total repeat content was measured. **(B)** Data was $\log_{10}$ transformed and an arbitrary cut-off for putative heterochromatic regions was designated at the point where the frequency of low repeat content scaffolds (red) is grater than 1% and high repeat content scaffolds (blue) is less than 1% frequency. **(C)** The density of NCBI RefSeq (green) and GNOMON ab initio genes (yellow) is super-imposed on the repeat content, revealing enrichment of non-RefSeq supported genes in the high repeat content regions.

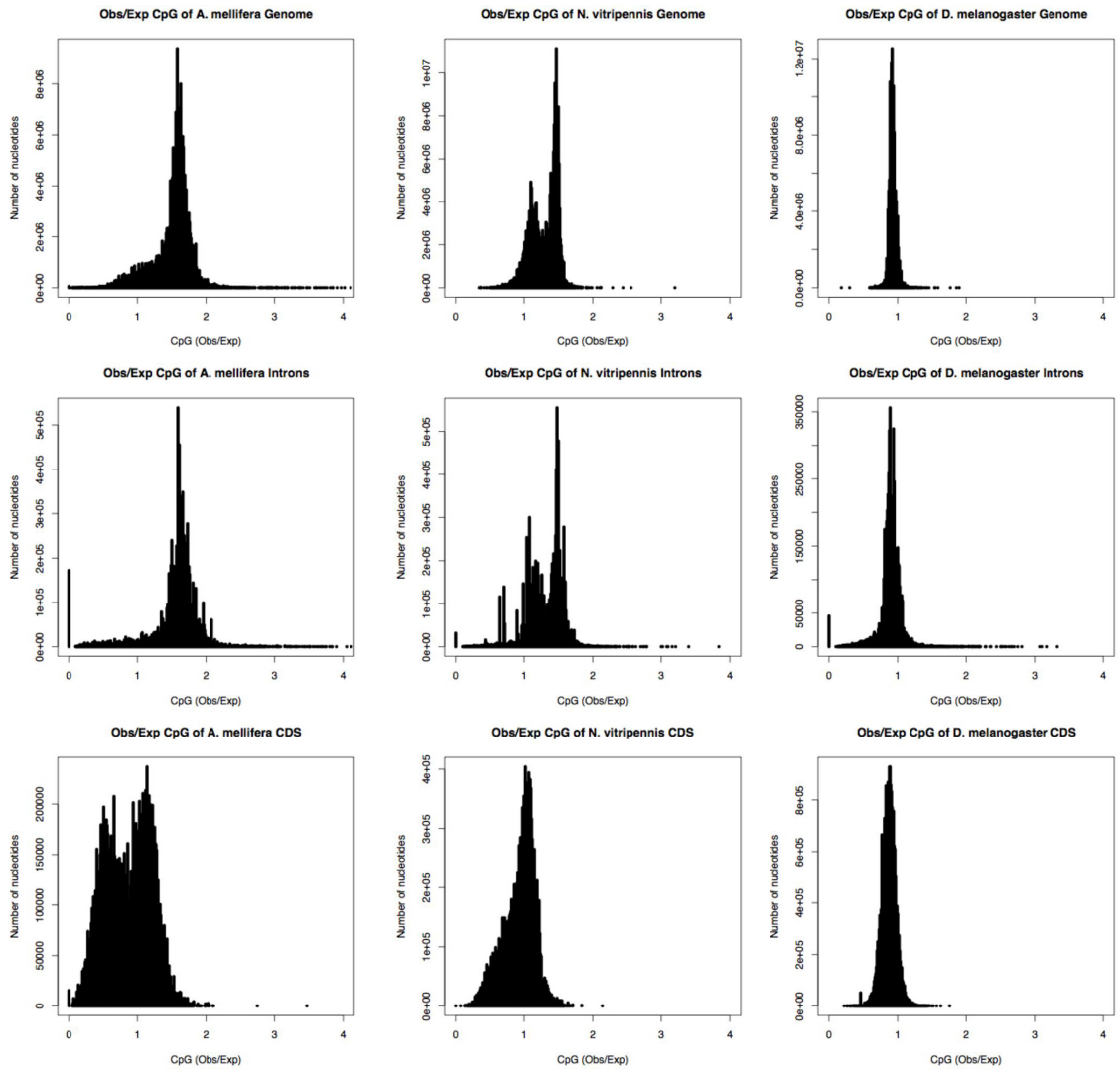


Figure S2: Distribution of observed to expected CpG ratios in genomes, introns and coding exons for *A. mellifera* (right), *N. vitripennis* (middle) and *D. melanogaster* (right). Number of nucleotides in genome GC content domains (top row), introns (middle row) and coding exons (bottom row) are plotted versus Obs/Exp CpG ratios, which were computed for individual sequences. Intron sequences from each gene and coding sequences from each gene were concatenated to determine average intronic and coding exon Obs/Exp CpG for each gene.

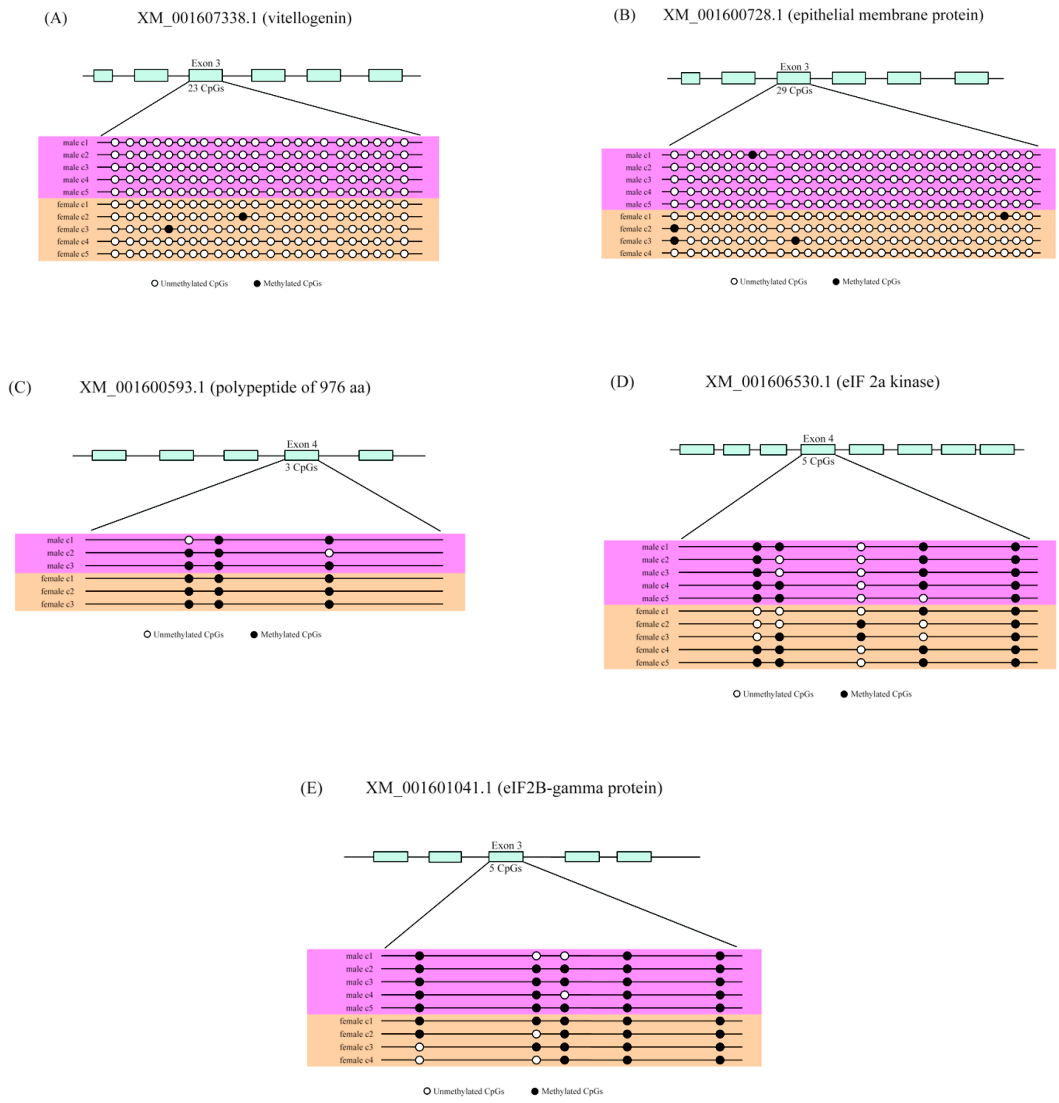


Figure S3: Distribution of methylated and unmethylated CpGs in five *N. vitripennis* genes as determined by bisulfite sequencing. **(A)** methylation profile of the XM_001607338.1 locus (vitellogenin). **(B)** methylation profile of the XM_001600728.1 locus (epithelial membrane protein). **(C)** methylation profile of the XM_001600593.1 locus (polypeptide of 976 amino acids). **(D)** methylation profile of the XM_001606530.1 locus (eIF 2a kinase). **(E)** methylation profile of the XM_001601041.1 (eIF2B-gamma protein).

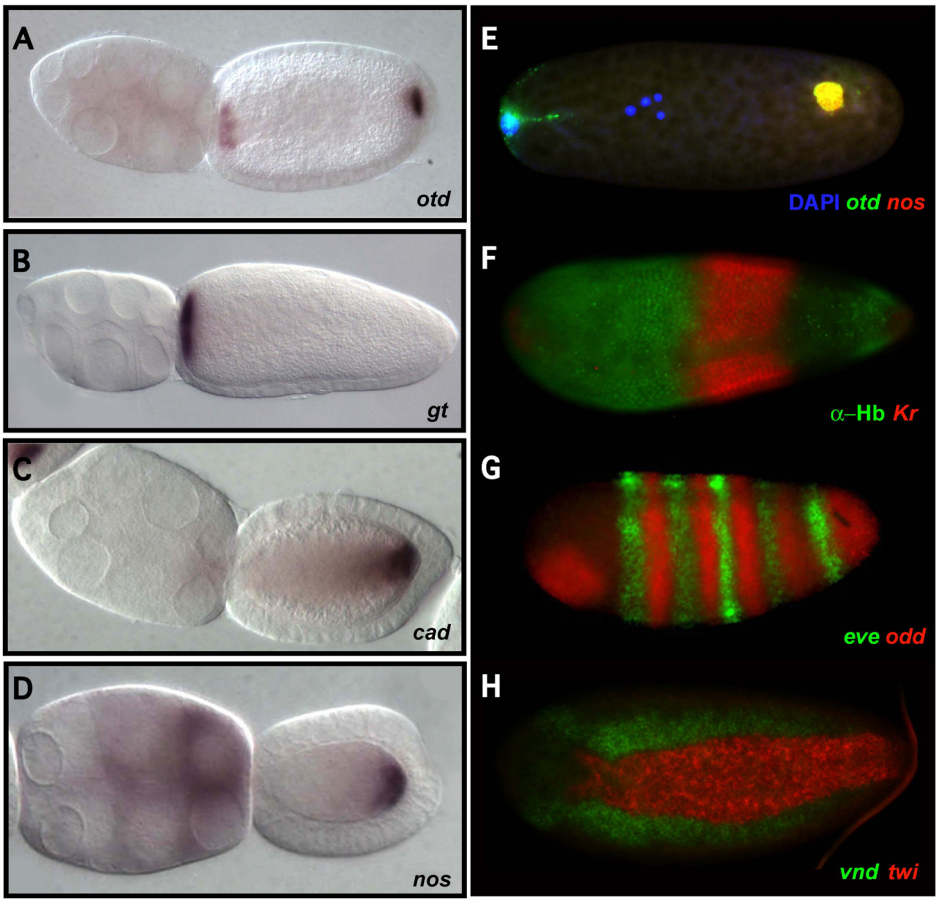


Figure S4: *Nasonia* segmentation genes. (A-D): maternally localized mRNAs in egg chambers (anterior is left, posterior is right). *orthodenticle-1* mRNA localized to both poles of the oocyte (A) carries instructive patterning information in the embryo. Anteriorly localized *giant* mRNA (B) represses trunk development, while posteriorly localized *caudal* mRNA. (C) restricts *caudal* function to posterior regions, and *nanos* (D) shows posterior patterning. (E-H): Extensive similarity to *Drosophila*, exist for segmentation genes. (E) From very early stages embryos (4 nuclei, Blue), *orthodentile-1* mRNA (green) is anteriorly localized. It is also co-localized at the posterior with *nanos* mRNA where it is associated with germ plasm. Expression of gap genes such as Hunchback and *Krüppel* (F), is largely conserved. (G) Pair rule genes *odd-skipped* (red) and *even-skipped* (green) are expressed in stripes in the pre-cellular blastoderm, like their fly counterparts. (H) The expression of mesodermal marker *twist* and neuroectodermal gene *vnd* indicates that the dorso-ventral fatemap of *Nasonia* is also similar to that of *Drosophila*.

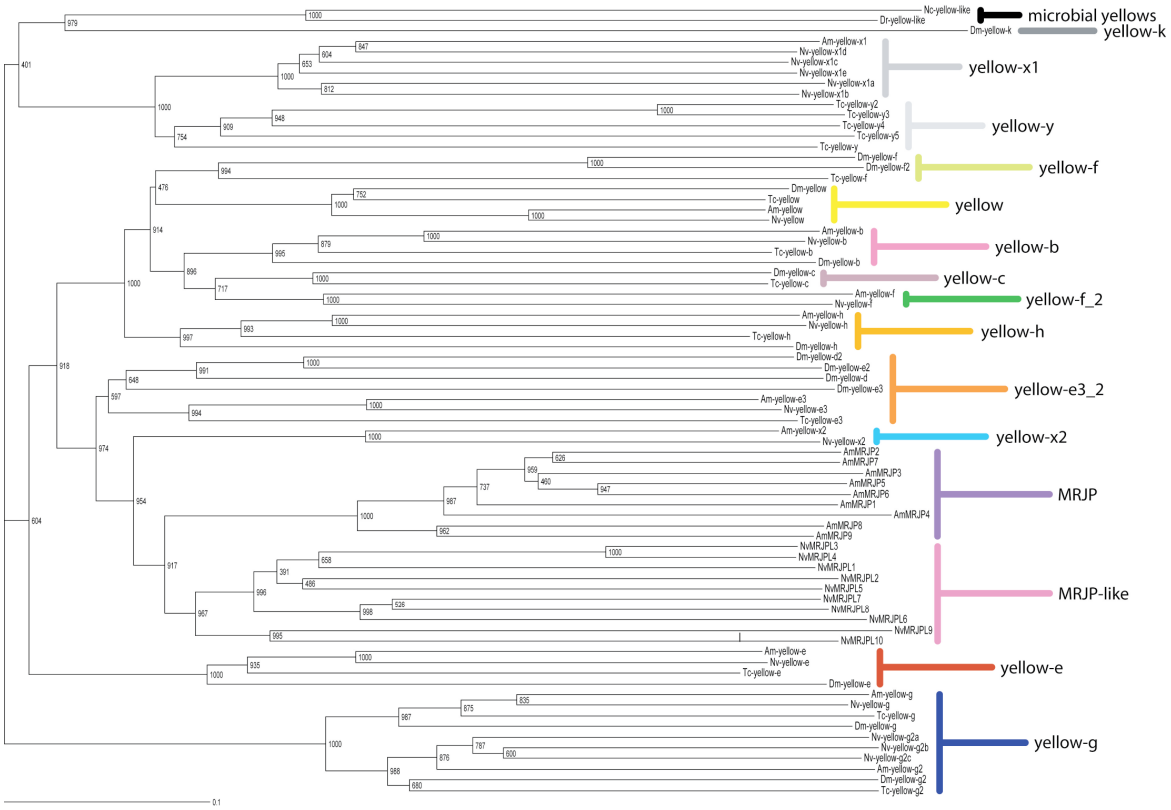


Figure S5: Simplified phylogenetic tree illustrating the relationship between the Yellow/Royal Jelly proteins in microbes, *Drosophila melanogaster* (= Dm), *Tribolium castaneum* (= Tm), *Nasonia vitripennis* (= Nv), and *Apis mellifera* (= Am).

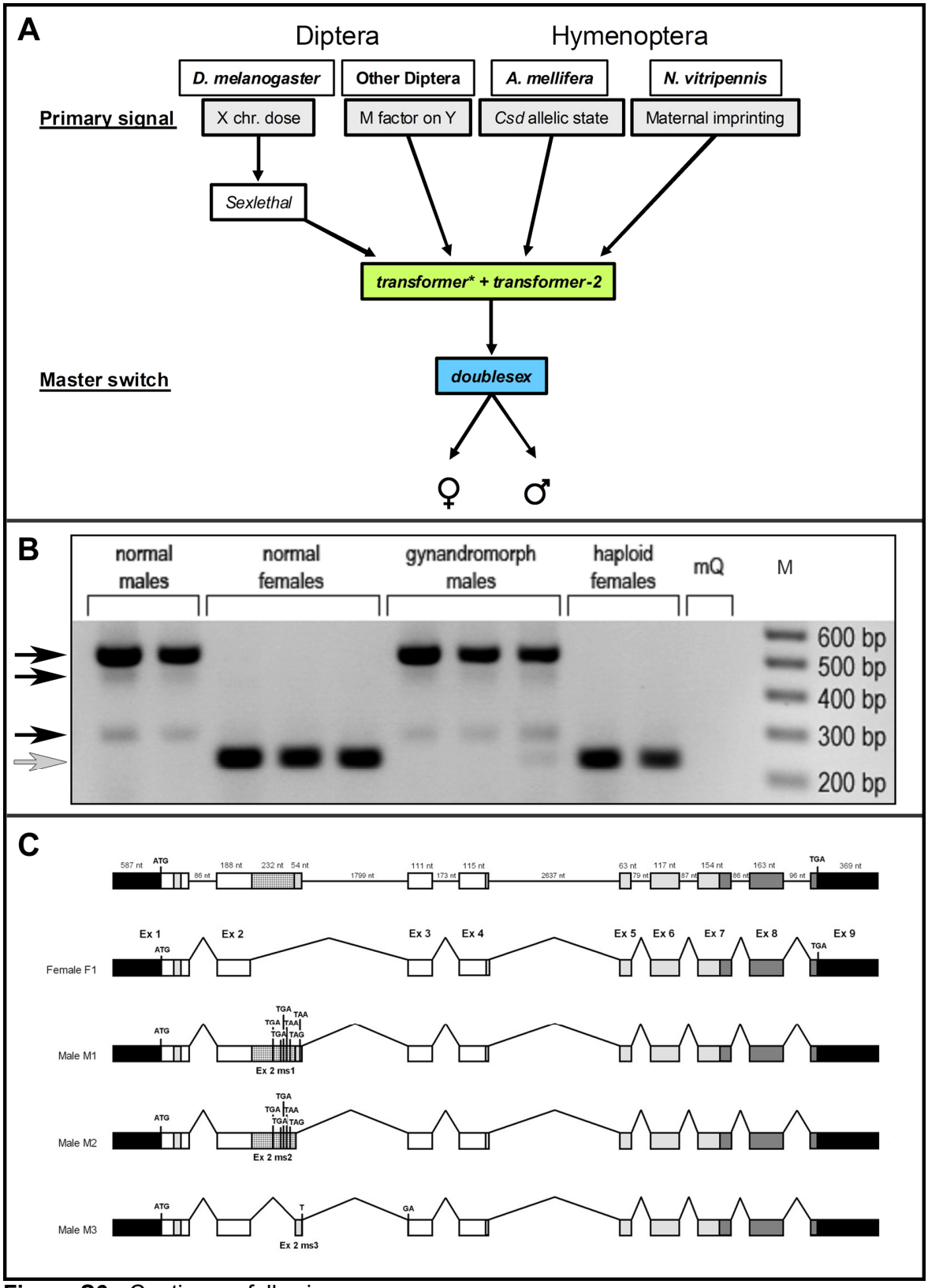


Figure S6: Caption on following page

Figure S6: **(A)** The sex determination cascade of Diptera and Hymenoptera. Diptera include *Drosophila melanogaster* (*S196-S199*) *Musca domestica* (*S200*); *Ceratitis capitata* (*S201, S202*) *Bactrocera oleae* (*S203, S204*) and *Anastrepha oblique* (*S205, S206*). Hymenoptera include *Apis mellifera* (S*188, S207, S208*) and *Nasonia vitripennis* (*S184*). *Doublesex*, *transformer* (* called *feminizer* in *Apis*) and *transformer-2* are functionally conserved, the difference between species is at the primary signal level. **(B)** Male and female spliceforms of Nasonia *tra*. RT-PCR of males, females and gynandromorph RNA. Black arrows indicate male specific and grey arrow indicates female specific splice fragments. Lane 1-2: adult male; lane 3-5: adult female; lanes 6-8: adult HiCD12 haploid gynandromorph, morphologically male; lane 9 -10: adult HiCD12 haploid gynandromorph, morphologically female; lane 11: milliQ; M: 100bp Molecular marker. **(C)** Structure of the *Nvtra* gene. In black: 5' and 3' UTR, white: coding region. In exons 1 and 4-7 RS-rich domains are depicted as light grey. In exon 7-9 a proline rich domain is indicated as dark grey. The shaded regions in exon 2 exons are male specific splice products and either contain or result in premature in frame stop codons.

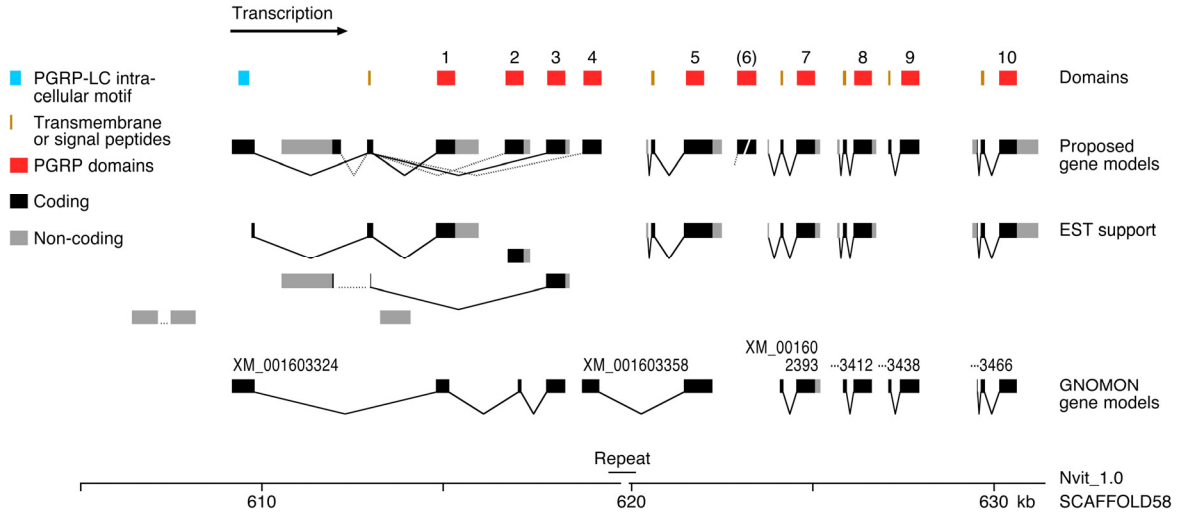


Figure S7: The major PGRP-LC-like locus in *N. vitripennis* encodes ten PGRP domains. Domains 1-4 are likely alternative splice forms of a PGRP-LC-like gene, with a shared transmembrane region and an intracellular domain that has sequence similarity to intracellular motifs in *Drosophila* PGRP-LC, LA and LE. One EST may encode an alternative intracellular domain, without conserved sequence motifs. Domains 5 and 7-10 are spliced to short exons that encode hydrophobic amino acids, which may either be export signals or transmembrane anchors. The SignalP and TMHMM prediction algorithms (*S209, S210*) give support for both interpretations, with some preference for transmembrane segments. Domain 6 contains a frameshift and is likely to be a pseudogene. Solid lines in the gene models indicate splice junctions with support in EST sequences; broken lines are predicted splice junctions. 5' and 3' EST sequences from the same cDNA are connected with a broken horizontal line. The GNOMON predictions (bottom) give plausible gene models for domains 7-10, but are not realistic for the alternatively spliced domains 1-4.

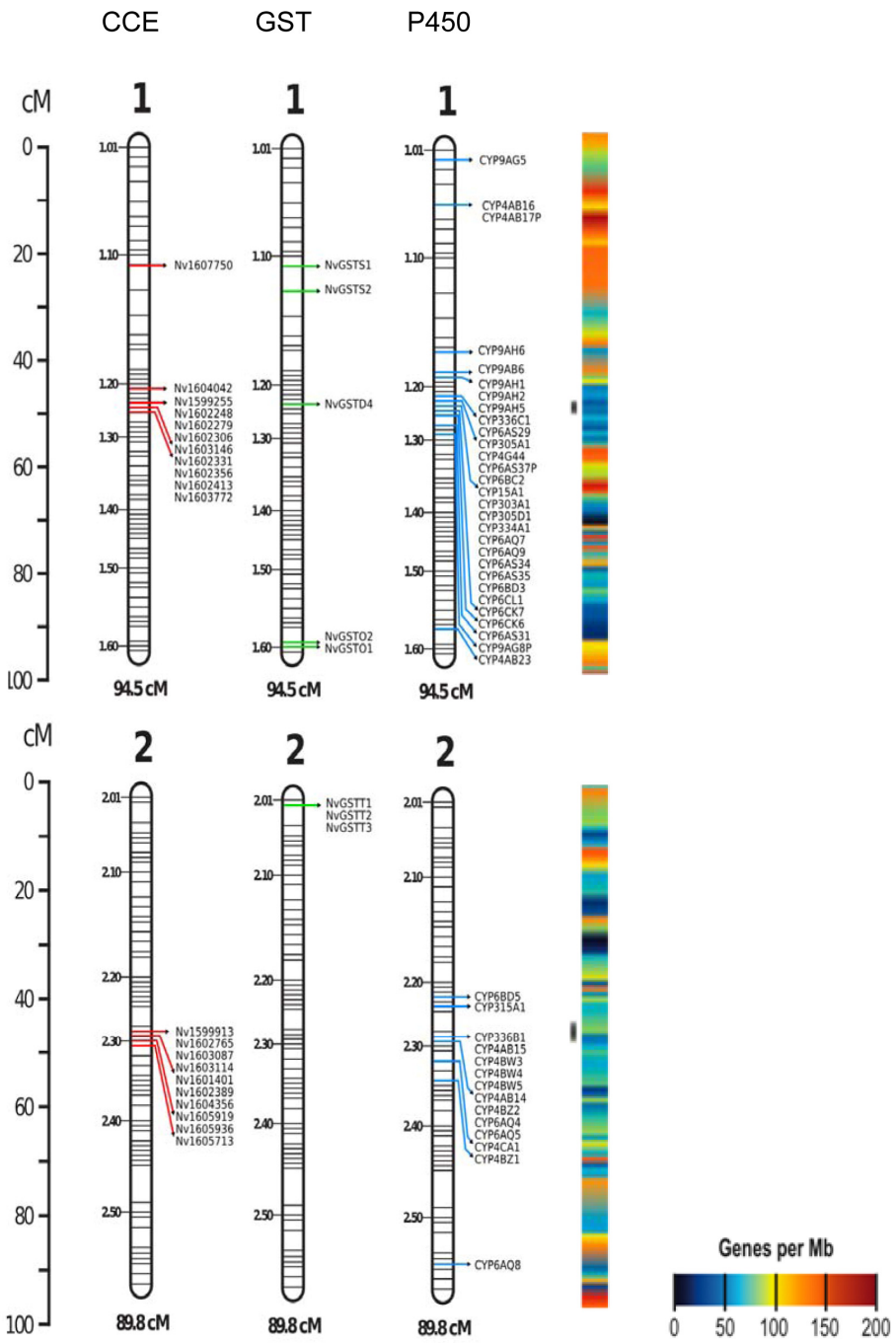


Figure S8: Detoxification genes mapped onto chromosome 1 and 2 of the *Nasonia* parasitoid wasp. Approximately 33% of all carboxyl/cholinesterase (CCE; red) glutathione S-transferase (GST; green) and cytochrome P450 (blue) genes are localized to chromosomal regions associated with low recombination (gray), high gene density and abundance of retro transposons (see Table S45).

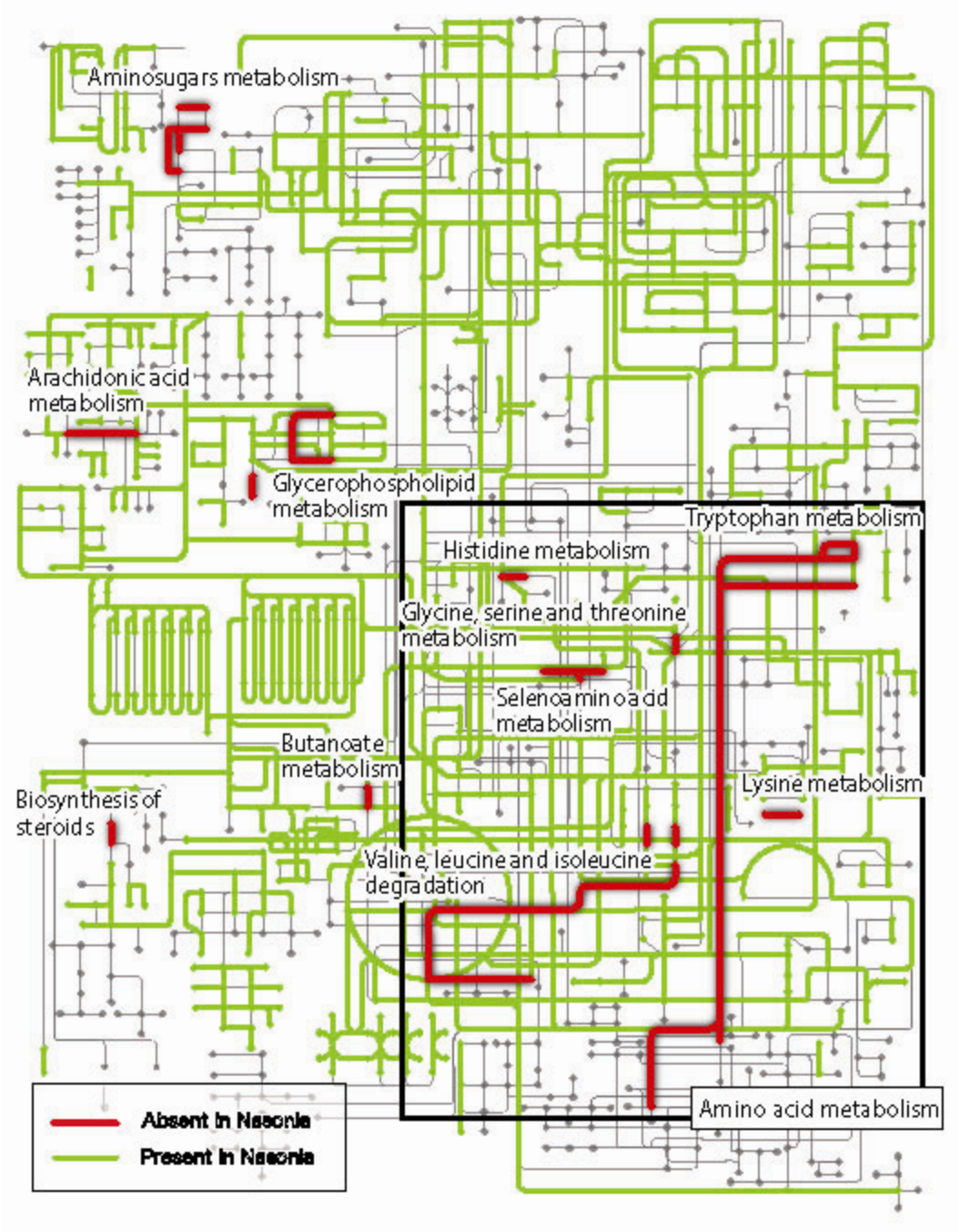


Figure S9. Kegg diagram showing functional pathways and clusters of gene losses in *Nasonia.* Genes involved in amino acid and aminosugar metabolism have been lost (or highly diverged) in *Nasonia*, possibly reflecting the specialized carnivorous diets of parasitoids. Gray lines represent pathways not found in the considered species set.

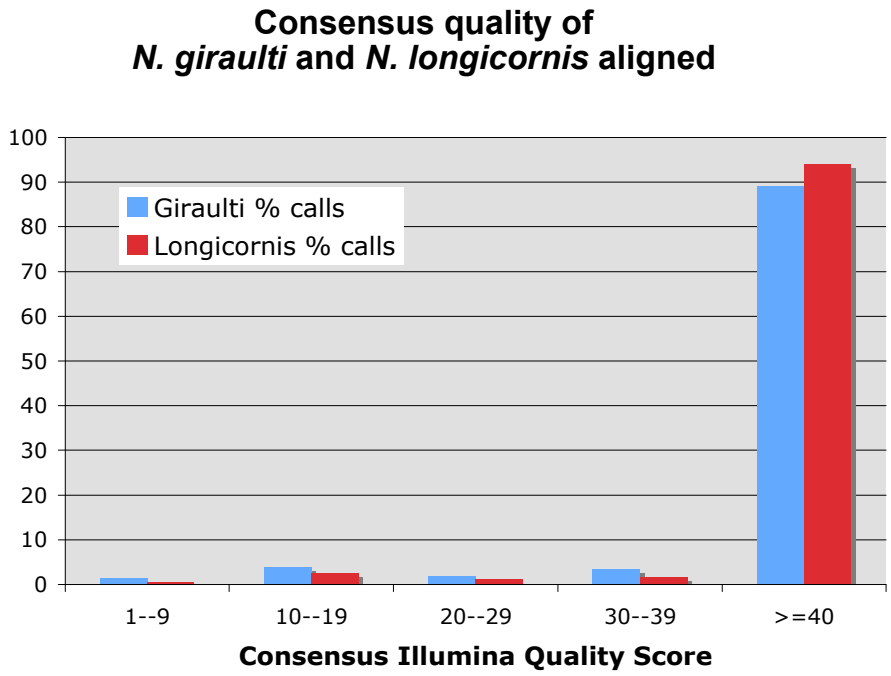


Figure S10: Consensus qualities of *N. giraulti* and *N. longicornis* aligned sequences. Note that consensus quality scores were capped at 90. 71.6% and 82.3% of *N. giraulti* and *N. longicornis* aligned bases respectively were capped at a consensus quality score of 90.

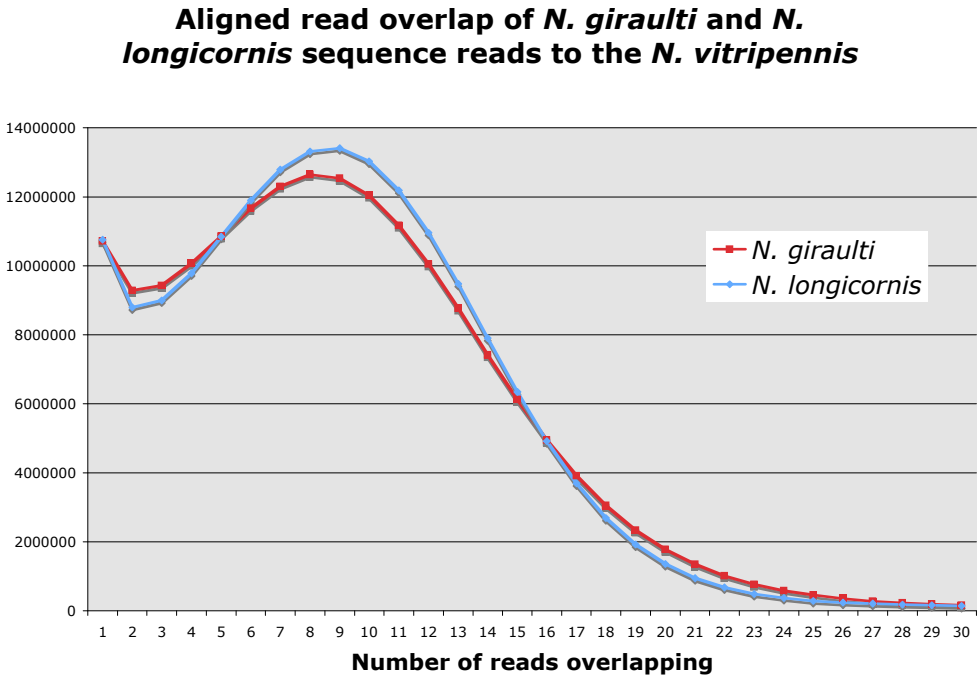


Figure S11: Read coverage statistics for *N. giraulti* and *N. longicornis* aligned sequences. The mode of coverage is 8 for *N. giraulti* and 9 for *N. longicornis*.

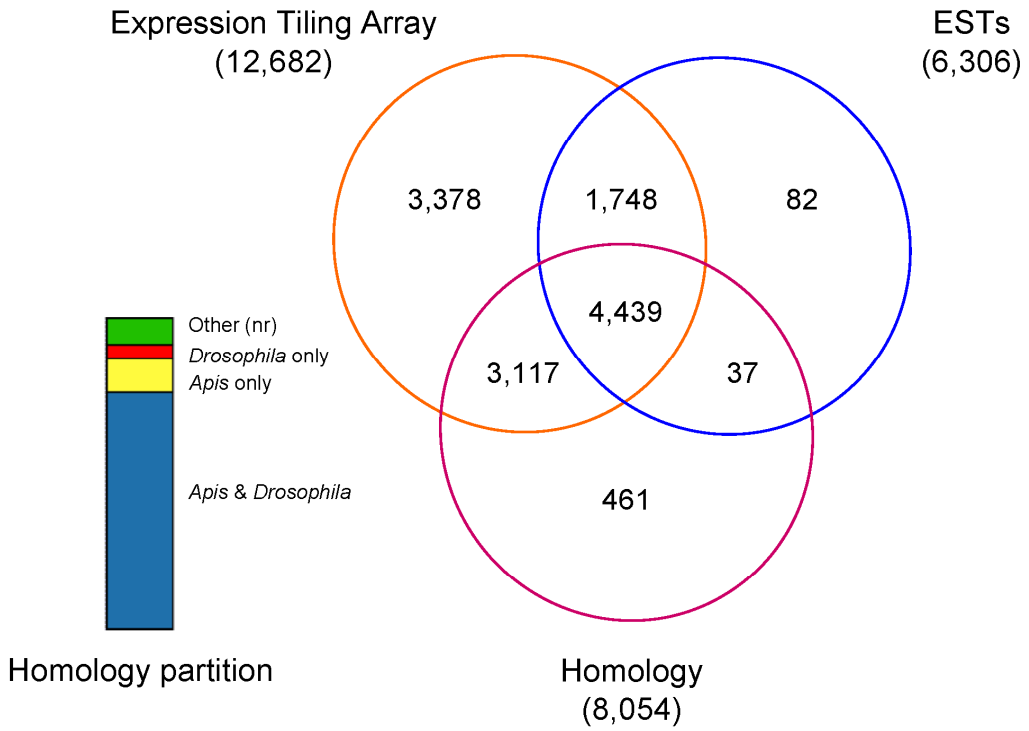


Figure S12: Venn diagram depicting 12,811 genes that are validated by expression data from expression tiling arrays and from the large EST sequencing project. Of the 8,054 genes with homology to proteins found in *Apis mellifera*, *Drosophila melanogaster* or other proteomes in the NCBI non-redundant database ($p < 1 \times 10\text{-}30$), only 461 genes with significant sequence homology to other proteomes are without transcript data.

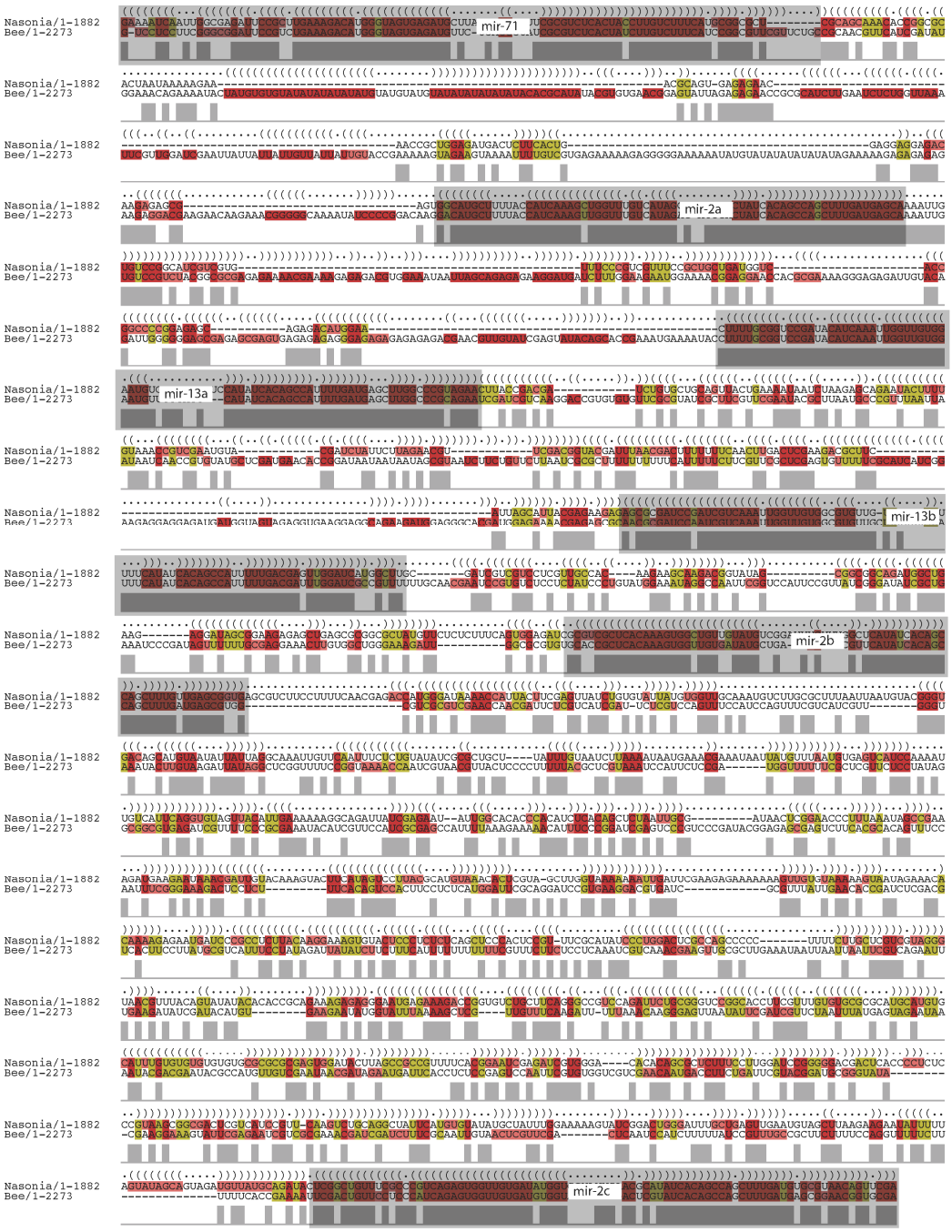


Figure S13: Genomic cluster of microRNA genes conserved between *Nasonia* and honeybee. A conserved cluster of 6 miRNAs (nvit-mir-71; nvit-mir-2a, nvit-mir-13a; nvit-mir-13b; nvit-mir-2b; nvit-mir-2c) spans a 1.9 kb in Nasonia, and 2.7 kb in honeybee. Apart from the miRNA encoding regions the alignment shows no significant sequence conservation between these two Hymenoptera species along the syntenic block. The region was color coded to visualize consistent and compensatory base changes supporting the common RNA secondary structure by folding with RNAalifold software.

(A)

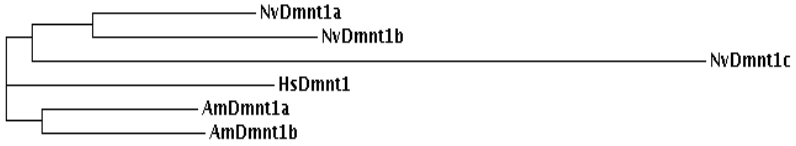
ATAGTTGATTTTGATTAGTGATCAGTGTGTCTTCGTGAGCCGCAGACATCGGCATCAGGCGTTGCTCACAACGAGATTTTTGTTCTCACCTTGCG
TTTTTTTTGTCAGCAAATATACTTATAGGTATTTTAAACAAGAAAAGGTCAATACACAAAAATCACATTGGATTTGGATGAAAAACATTTGAATGTAT
ACAAGTATATATGAAGATTGAgtatgtaattttattatactatcatattttatataatttgaagacacttggaaattgtgttcagcaagattatatgaatttttttttaacactgaaaggattttaagtttttacattatac
ttatattgtcaaaaacctggatgaatatgcatgcgattttaaagatttttttgcattcctgacttgtttgactagaaaaatcgtgatgataaaaagtaatttacatttacacatatacacatgcaattatagaaaagatttttttaag
tagaacgcatctatgtaacttttttgttttgtcaaactcttaagagaacctgcaatatctaacaatatttcataaaatataaaagtatattagggttgatttagttttatttatttaataaattaaaaaaagagttgtataaaatattt
atttttccttaatgcataacgtctttgttagctaatacttatacatactaaatacgcgtaatatcactgatctcatcaaaaagttttaaatgaagaaatttaaagatgtatcttgcattacttcctctcaacatgtatacatgtcttta
ctgaaactgtaaagtactgtttgaccactttctttagtaatttatggtttcaaggcaaaatggtacattgactcaagtagaacaaagaaaatgaaaatgttcaattttctgtgttgttagaatctaaatctcaaattattttaatgt
gatgatttcttcacattcaccgttcacaaaatcttaggtttcttaagaaaaaaattaagaattgtttattttaaatctttttaagtgattatcacactgctaaataatagttaccattctgcgtattgtcaactaagatattattagtc
agaccatgtgctgtgttatcgatatcgtcagtacaaatttctagaaaatctgtaatactttgttcaggttcttcttggttttgagcattcaatctttacaccttgcttggtcacttttttgctttacataaattttaggaaagatcgttgg
gttgtacgaagtttttcttggatcatttgatttttttccctccttcaaaatgtgcactacaaactctagaactctctgttggtctccagtcctcatccggatttctataaataagtataaaataaataataatttgccaacgctccaag
caatcatttcaaagtttagatattcatttaatatttttattaaagttgaatgagacttaacctaataattacagggtgatttcgtttaagaattatttaaactactaaaacttttacatatttcattacaaaataaagacaatcctacaa
gatttgaaaaatcaacaaaccccttgcacaagtatacttacgtttcacttccgcgacgagcggctaaaagccatatcatttttttttcagcttcataaggctgaaagggaaacctatagaagtggagatcccgattattgcg
ttctgagttattgcaccatactacacagcatttacgatatgcaccgcgttttttcgctacatttttttattttttcctaactaaaatttttgacacttgctgcagaacttataatattttacgtcttctgctacaaaattgtatatatgttttgct
aaatgtatccgcgtggcggagcagtccccgatacgtttagcctttaggatatggcaaatataactctcggtcgtgacgtcatcaacatggacgcccaggccatgaaaaatacattagtagtatacaaagcaaagcagt
ttttactaatttttgacagaaaaataacgtgacagctcttaaatttatatatgtaataacgcatataatttactataagatgatggtactttttccctccctgaaattgttaatatgttttatttaggatgatcaaagcattgtataggt
tatagaggtaagaccacaattaagtatttgtagctagaaatagtaaaaattgataaatattaataaaaaattattgtatattcaagcaacatggaatgtaataaccaaaaagaggcaggcatgcaattttcacaaaaa
aaaccagtagtagaacatattgttgtgtttctggatgcactagtaatgctataaaaaatcctgttacgttttcactttgccctgctttacatgagcaattaaaactaataattgagccatttattccaataactccactaatttcat
gaggtggtcctctcaaattggatatttgtaaacagtatgcgcgtatatttaaattctctgccgattctaatttaccacatttcaatatgtgcttagaattaagcactctttctccttttaataaaaatttctacttgagcttgtacactc
aacaaattaaaaaatttctcgaagtgttactatcattgtttatcgtgagaggttatgctccaaaacgagtttttctattattaaagaaaatacaatactgtcacgcaaaaaaaaaaaattaatctacgtattttctgatgttatatt
aatgcaccaaaaaaactattactctatagttcactgctttacgtaaaaaattagaaaatactgctctactccgtatactgacaatgtattctgcatggcctgggtgtccatgttgatgacgtcacgaccgagagttatattg
ccatatgcttaaatatgctttaacgatgtattgagattattatatttcttagcgtaaatgggacatatatttgagtaaatcttaaattcttctcgttaacttgatctgattagagaatactctttacttatttcatatgcttatctgttcaat
gatgcactgaatcgtttctcgcttataaaatttttatatccgacagttgtttttttactaagttaccaagaatttttaaatagaaaatttttcttacaagctttttatataaaacgtgcacttatttgattaaaaactgataaatatttaatcct
actctttctgaaaatacttgaaatacaatgctatttcgtccaatcgttaaaagcgatttttattggattcatgacatgctcagacaaatagcaaatacaaaaaaaaaaattcaacattttgaaaattgtgtcaacgtaaaag
acaaaccacgtaaatttattcttcgaaaaatgtatggtataacaattataccgatttaatgctcaacgaaatgtggaattgccgatttttttcagcgttaaatatagatcaaatttttgtcttttctacttatatatatatatatatc
acagctatctatgtatatgtatatatttttaggcataaacttagttgataatcttgaagtcggagtatcaaagtatagcttaaactactaaaatcaagtaatatttatacagaattcgtaaaatttaaataggttattcagcaaa
ttcaatttataggtcgcgacccgctaccgacgtcacagtgtataacacttttatttgccatcggggcgctgccgcgcaccttcatttttcgctattatacacagtttttatttctaatccttaaaatatgcaaatttatgacatctttt
tttattccagGCGACTCTCTTTTATATAAATGCTGATTGTGTTGAATTGGTTTAAACACTAAATTACAATGGTTCCAATGGATTTGTCCATAACTTTTCCAA
CCGGACCTGAATACAGACAATCACTGAGTGATTTAAAACAGTTACTGTTCACTGAAAAAATGGGAAACTCTGACCAGATTTTGGCCAGCATTAAAA
ACGCGGATGTAGGACATGAAGCTAATCTTCGTTTGCAATTCTTAGAAAAGCGCGAAGTCCCTACTTCAGAAGTTAATAATATTTTTTATGAGTTTC
AAAAGAGACACAAATCAATGAATTACAATGATCTATTCAAGCAAAAAGACTCAGCTTCCAGTTTCTTTAAAAGAAGAAAAAGAACAAAAGGTGAAA
CCGTTTATCGCGAGTCGAAACCTCTCGTCAAAAATCAAGCTCATCCTGCACTAATTCGCGAGAGCTATGCAACAATTATGCACCGAGTAATGACT
CCTAAAATTATTGGGAACAATCGAAACAAGAAAAAGATCTTGAAAAGTTTATTTAGATTCATGCAAACCGCAAAAACAGAATCGTTCTCCTTGCTTC
CTTTAGTAAAACAAATCAATgtaattttgtttctcatttttcaataattgatgtgcatgatctattataatagcacatattaatttttttcttatcattattttcttgctgaaattcgaagATTGAATCTATTTTATG
GCTCACCATCGAAGACCAAGAATTGAAATTGATTTTTGTTGCTAAACTCATGAAGTGGATGTTTCGATTTTTTGTACTCCTCATATGCAGGTACTTC
ATTATTACATACGTTAAAGAAGAGAGAGTTTTCTGTTTCATCAAAGAATTAAACAGAGACATTAAAAGAGATTTAGCAACAAAGATTAAGAAGGGC
CTGCTGGAAGAATGTCCAGAAGATAGTATTAATTCTACTGAAACAAGAATGTACATGGTGCCAAAAGCTGATGGTTTTAGAACTATCACTAGACGA
AAgtacgtatgtacaaaagtagaaatacgttttactcatacggatatgtatatgtgcaaatccatgatggagtggaaacgagttagatgctaaaagctgaaaatgtaaattggagggcattccaacttactagatcatt
ttcttcggttctttaaaccatttaatgcttcaaaatgcgcgcgtacatcgccgattgcttgccaaacttttttgcaggacagtttaaattctcttatgaggatttgcacatatatgttaatgtgcgataataaacgttttaatatgcttt
ttagATATCCTGCTGATGAACTTAAAAGTCTAAGAAAATCTTTACGTGTTTTGGATTTTTTGTTTGATAAATATTTGAAAACTCCTGATAGCGTATTTAA
TGATGCTTACGTCGACGTTACAATCGAAGCAAATGAATCATCGCAAAAATATTACGTTACGGTGACTGACATCCGCGATGCATTTGGATCGATTAT
ACAGAgtaaataagtcgtattatcgacgccagctatataattatgagaaactttcaaatcctaaaaatgtatatgacaatgacaatatatttatttgtagATAAACTGTTTGATATTTTGACAGCGTT
ATTCGAGCTTTTGCCGGAAATATTAGAATTTACTGCATTCTATGGTGGTAAAAACTACAAATACGGTTTCCCACAGTTTATTAATGGCAAAGGTGG
AAAAAGAGTCACCGCACAGAAACGCAACATTTTGGCAATGATAGAAAAAACTATTTTTAATACGAAGgtaatttatatttgtgatttatctataataattaaatagatttttttc
atcaaaacaaatcattcgctctcatagGTCGCAATGCACAACAAACTGTATAACGTTTGCAAGGGTGTTACGCAAGGCTCAATTTTGTCGTCAAAGTTAAGCGA
TATTTATTATTTATCGATGCAGCGAAAATACTTCGCTGATTTCCGAAAATCTGGCCGGATATTTAGTCATGTGGACGACATCATATATATCACAGC
GAATGAAGAGTTTGCAAAACAgtaagttaaatacatgcaagtgtcaagagtgtcactgaagacacgcatatcgtactggtatttaccttttttaattggatgaaattctaatgaaaggtttattgatttttgtttt
agATTCTTAAATCGCGTAGAAAAGGGTATTCCCGAGTACAATTGCTGGTTTAAGAAATCAAAAATTGTGTCCAACTTTAGCAGAGATGTTGCTGTTT
TTGGTGAGAGTCTGGAAAACACAATCAAATTTTTGGGCTGGTCGATTGACGGTACAACGTTGGAAGCAGAGCCATATGTAAAATGCTCTATCCGG
CACTCTATCTCGTATTCTGCATCGTTAGATATTATGCACgtataattctatgttactcatacatatctgcatataaattatatacaaaataactatttaaattgaatttgtgtgcatgcatagaaaat
taaagAATAGACTTAAAAACCTGGAGTCAATAAAAGTAAAGGAGCCAGCCTTAAGCAAATATATTAACAGCAACAACAAAATTTTAAGTAACATAAAT
GTAATTGCCAGATTGCAAGCACAACGGTGCTTTGCAATCGCAAACAGCCTCAGTGATACTACTATTTTGACTTTGGATGCCATTGCTGAGACCAA
TAATAAAATTATGAAGAAAATCAAGAAAATCACCAACTGGAgtaagcaatttttagtttttaaataactaaatttaatttttattatagtagctagccttaaatacgacattacggtctgtgatgtc
atgacgacgtgaacttatcactttcaggcatttttcaagcccatgtacactgatgtaaacttaactatacaatgatacatttgtaaaatattcaaatttttttatagCTAAAACTGAACAGTTGGAATACAGC
AAAGAAATTCGACATATTTTTTGGAAGGCATATTACGCAACTTTTATTAAAAAAAGTGATCTTTTTCCAAACCTCATCAAGACGTTTAGAAAATTGAC
GAGAAAAAACAATGTGCACAAGAAGCGTAAAATTCGAAAATTCGTAAAAAGATCGTCATTAGTTTCCAAAAGGGCCAGACATATGTAA

(B)

MVPMDLSITFPTGPEYRQSLSDLKQLLFTEKMGNSDQILASIKNADVGHEANLRLQFLEKREVPTSEVNNIFYEFQKRHKSMNYNDLFKQKDSASSFFK
RRKRTKGETVYRESKPLVKNQAHPALIRESYATIMHRVMTPKIIGNNRNKKKILKSLFRFMQTAKTESFSLLPLVKQINIESILWLTIEDQELKLIFVAKLMK
WMFRFFVLLICRYFIITYVKEERVFCFIKELNRDIKRDLATKIKKGLLEECPEDSINSTETRMYMVPKADGFRTITRRKYPADELKSLRKSLRVLDFLFDKYL
KTPDSVFNDAYVDVTIEANESSQKYYVTVTDIRDAFGSIIQNKLFDILTALFELLPEILEFTAFYGGKNYKYGFPQFINGKGGKRVTAQKRNILAMIEKTIFN
TKVAMHNKLYNVCKGVTQGSILSSKLSDIYYLSMQRKYFADFRKSGRIFSHVDDIIYITANEEFAKQFLNRVEKGIPEYNCWFKKSKIVSNFSRDVAVFGE
SLENTIKFLGWSIDGTTLEAEPYVKCSIRHSISYSASLDIMHNRLKNLESIKVKEPALSKYINSNNKILSNINVIARLQAQRCFAIANSLSDTTILTLDAIAETN
NKIMKKIKKITNWTKTEQLEYSKEIRHIFWKAYYATFIKKSDLFPNLIKTFRKLTRKNNVHKKRKIRKFVKRSSLVSKRARHM

Figure S14: The *Nasonia* telomerase or TERT **(A)** gene and **(B)** protein. Introns are shown in lower case and exons in upper case. The apparent start codon is 59 bp into the second exon.

A



B

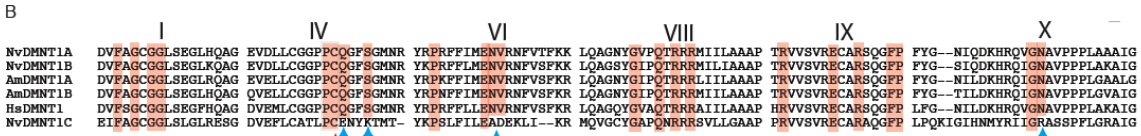


Figure S15: *Alignment of the catalytic domains of wasp, honeybee and human DNMT1 and DNMT3 orthologs*: **A.** Phylogram showing the evolutionary relationship of wasp, bee and human DNMT1 proteins. The branch lengths are proportional to the inferred evolutionary distance based on the number of differences between the sequences.

B. Alignment of DNMT1 orthologs showing six highly conserved motifs. The highlighted amino acids are highly conserved across prokaryotic and eukaryotic DNA (cytosine-5) methyltransferases; most of these amino acids are clustered around the active site of the enzyme. Although NvDNMT1C has the invariant cysteine (red arrow) required for catalytic activity, other highly conserved residues typically found in DNMT1 are not found in this protein (blue arrows). Sequence files used for this analysis: NvDNMT1A (XP_001605635.1), NvDNMT1B(XP_001600175.1), NvDNMT1C(XP_001607336.1), AmDNMT1A (XP_001122269), AmDNMT1B(GB30166), HsDNMT1 (NP_001370).

C. Alignment of DNMT3 orthologs showing five conserved motifs. Sequences used for this analysis: NvDNMT3(XP_001599223.1), HsDNMT3A(NP_783328), HsDNMT3B(NP_008823), and AmDNMT3 (Table S58, available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html).

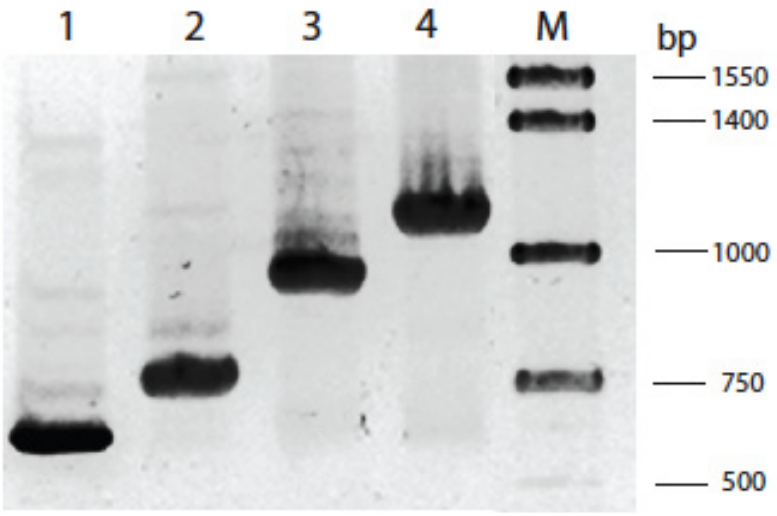


Figure S16: Visualization of NvDnmt1a RT-PCR products by ethidium bromide staining after agarose gel electrophoresis. The PCR products were generated by amplification of cDNA prepared from total RNA isolated from male pupae. Primer pairs span splice sites predicted by the RefSeq gene model. In lanes 1-4, the left primer was located in exon 4 and the right primers in exons 5, 6 7 and 8 respectively. For each primer pair, the size of the PCR product was consistent with the gene model. Lane 1: 670 bp; Lane 2: 789 bp; Lane 3: 980 bp and Lane 4: 1150 bp. The entire gene model was confirmed in this way.

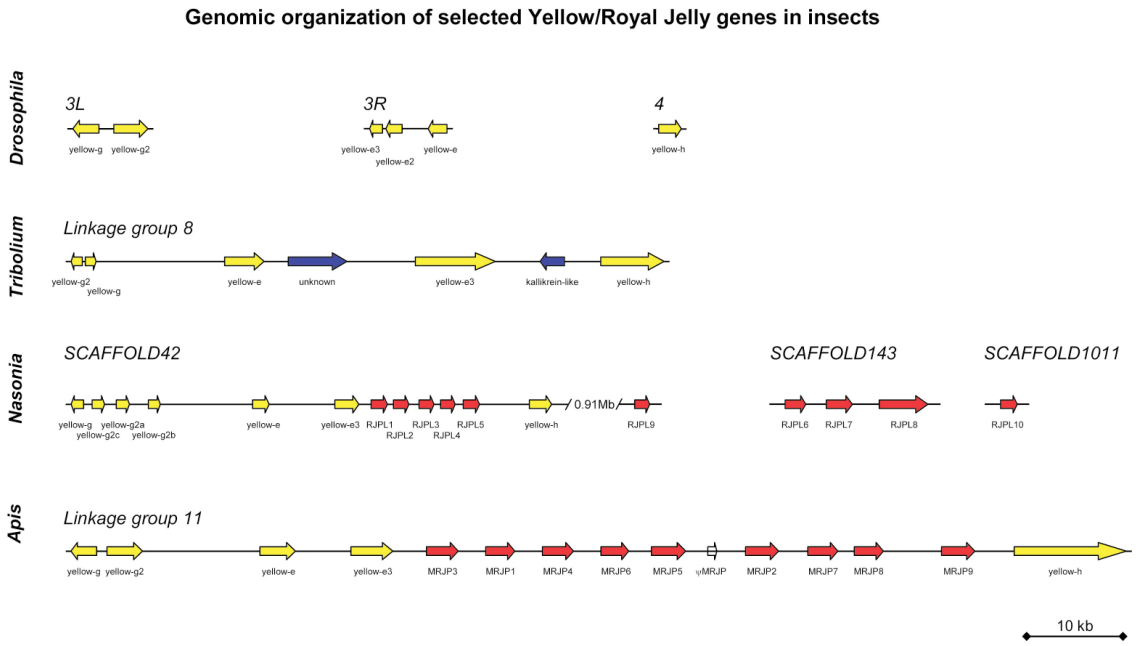


Figure S17: Genomic organization of selected Yellow/Royal Jelly genes in four insect species: *Drosophila melanogaster*, *Tribolium castaneum*, *Nasonia vitripennis*, and *Apis mellifera*.

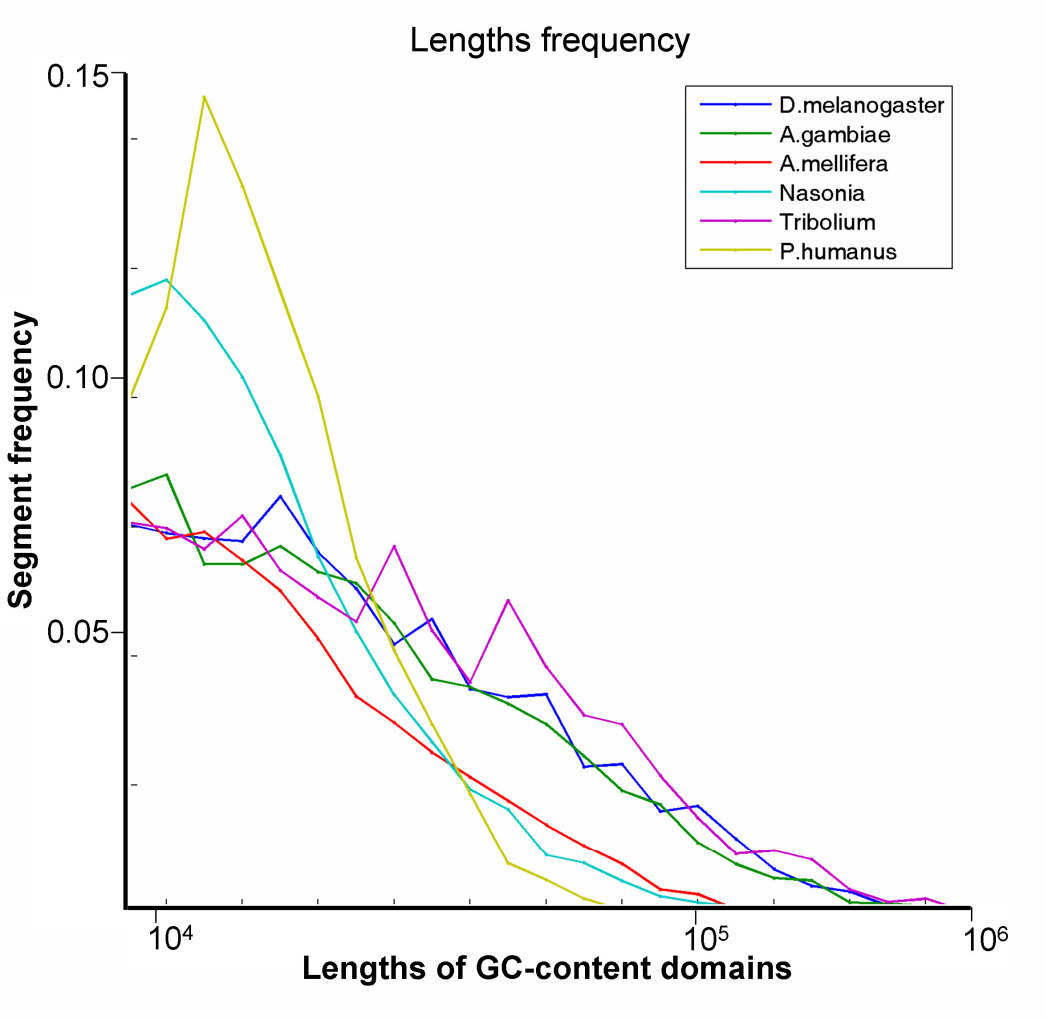


Figure S18: Frequency of homogeneous GC domains. The frequency of GC-content domain lengths of six species (three of which were recently sequenced). These plots illustrate the distribution of homogeneous GC content domains (isochores) of different species. It is noticeable that recently sequenced species have high abundance of small size GC-content domains (3-17 Kb) relative to the other insect genomes. This may be the due to the use of scaffolds instead of fully sequenced genomes (that were not available to us). It is interesting to note the similarity of the domain distributions of bee and fly that have very different GC content. These similarities and others indicate that GC composition may obey similar rules in all metazoans.

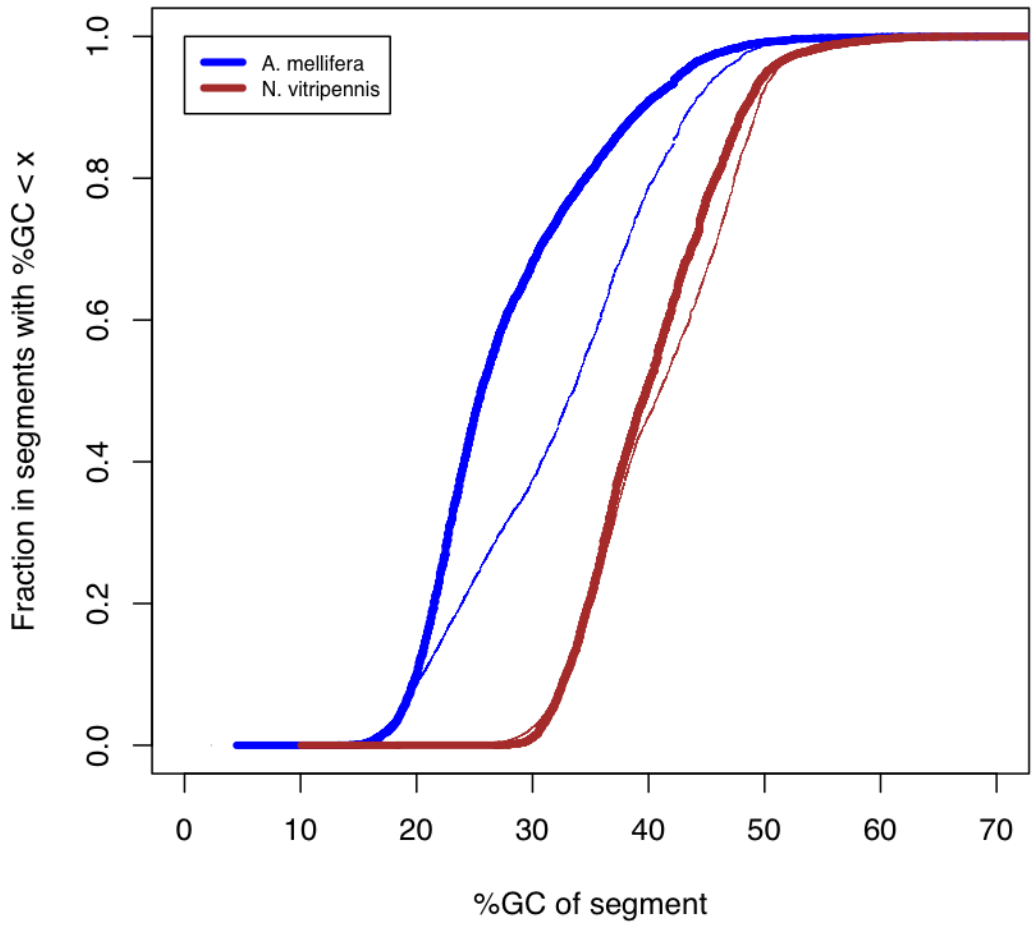


Figure S19: Cumulative distribution of the fraction of nucleotides with GC content in GC content domains containing genes (thick lines) and all GC content domains (thin lines) for *N. vitripennis* and *A. mellifera*. If there was no tendency for genes to occur in GC content domains of a particular length in a particular genome, the curves for the thick and thin lines in each species would be the same. The GC content domains in which *N. vitripennis* genes occur are slightly lower in %GC than all the GC content domains. The pattern is far more pronounced in *A. mellifera*.

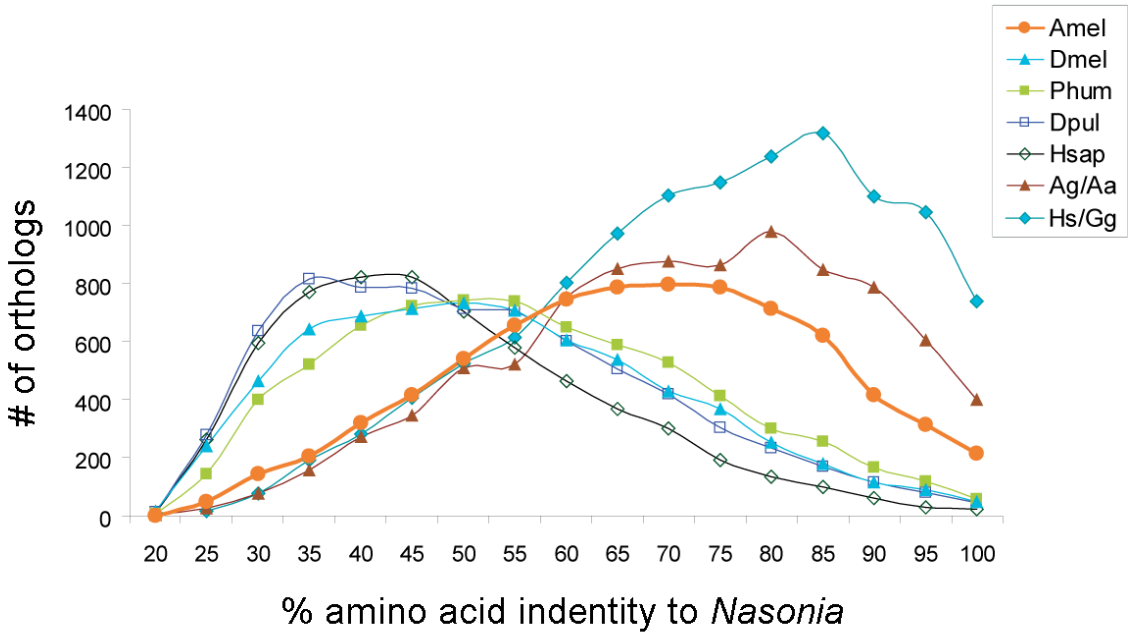


Figure S20: Distribution of amino acid identity of single-copy orthologs between *Nasonia* and the other species, *Apis mellifera* (Amel), *Drosophila melanogaster* (Dmel), *Pediculus humanus* (Phum), *Daphnia pulex* (Dpul) and *Homo sapiens* (Hsap). Note that Ag/Aa distribution is a comparison between *Anopheles gambiae* and *Aedes aegypti*, shown to highlight the similar level of molecular divergence as between *Nasonia* and *Apis* (Amel).

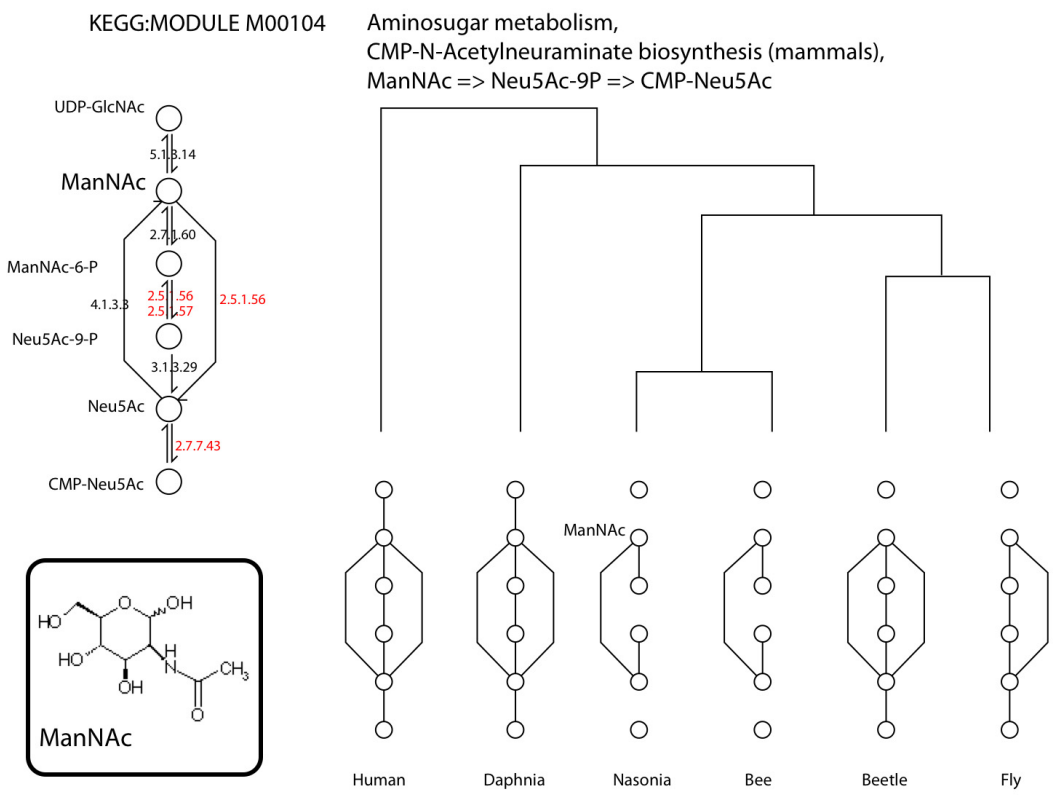


Figure S21: *Nasonia vitripennis* aminosugar metabolism. Enzyme numbers (EC) written in orange represent Hymenoptera-specific losses.

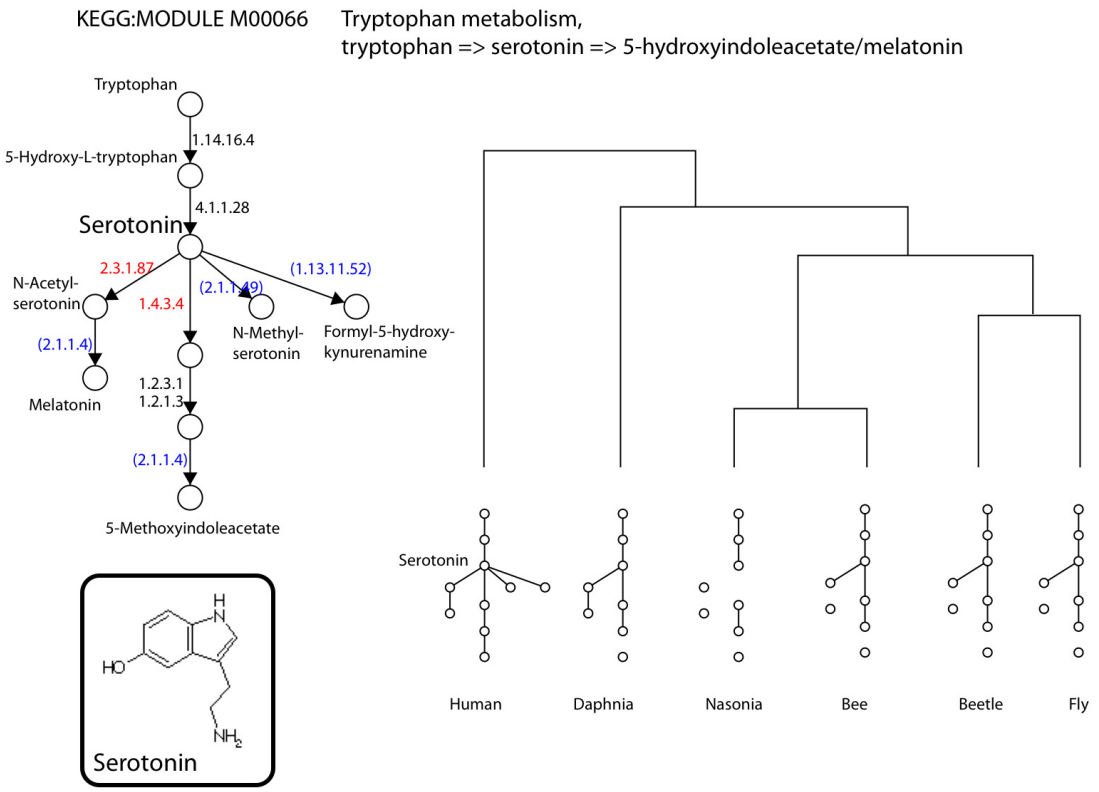


Figure S22: *Nasonia vitripennis* tryptophan metabolism. Enzyme numbers (EC) written in red represent *Nasonia* specific lost enzymes and those written in blue represent arthropod-specific lost enzymes.

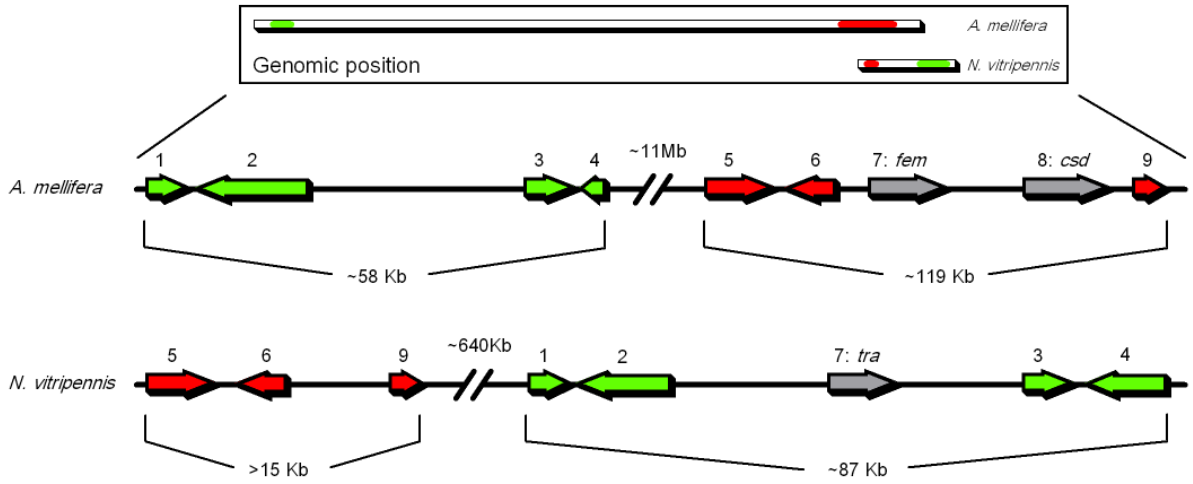


Figure S23: A schematic overview of the synteny between the SDL genomic region (*S188*) in *A. mellifera* and the genomic region of *Nvtra*. The numbers indicate similar genes between *A. mellifera* and *N. vitripennis*. Arrow 7 and 8 in *A. mellifera* indicate *fem* and *csd*, respectively, arrow 7 indicates *tra* in *N. vitripennis*. (Gene number followed by the NCBI gene accession number, 1: GB13000; 2: GB16089; 3: GB16151; 4: GB13465; 5: GB11211; 6: GB13727; 9: GB30480).

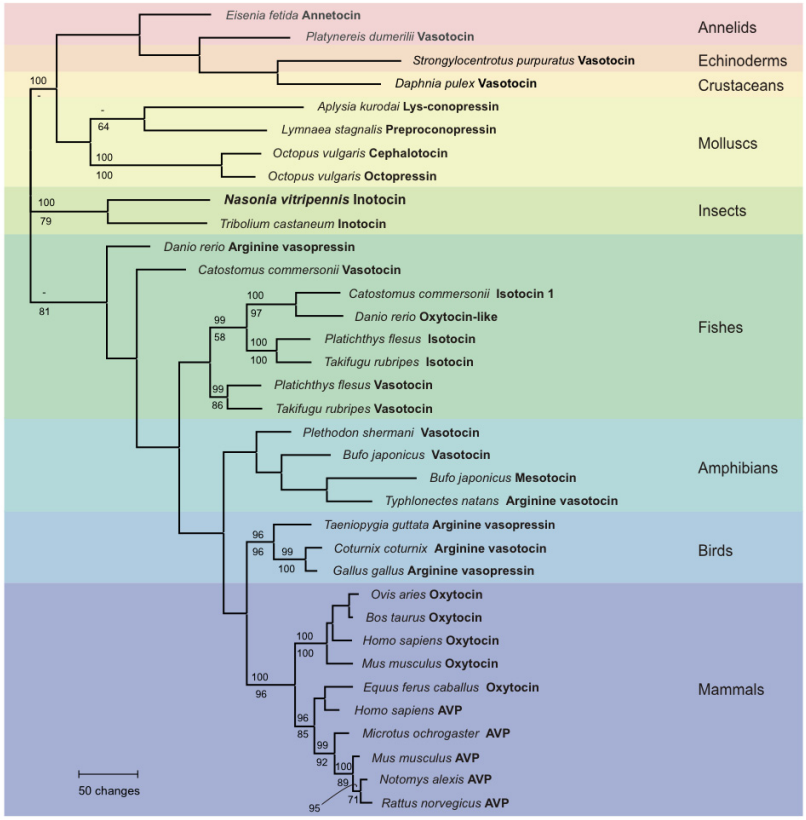


Figure S24: Phylogenetic tree of oxytocin/vasopressin-like genes. Maximum parsimony tree of oxytocin/vasopressin-like coding nucleotides from various taxa. Bayesian posterior probabilities and nonparametric bootstrap proportions (1,000 replicates) are shown above and below the branches, respectively.

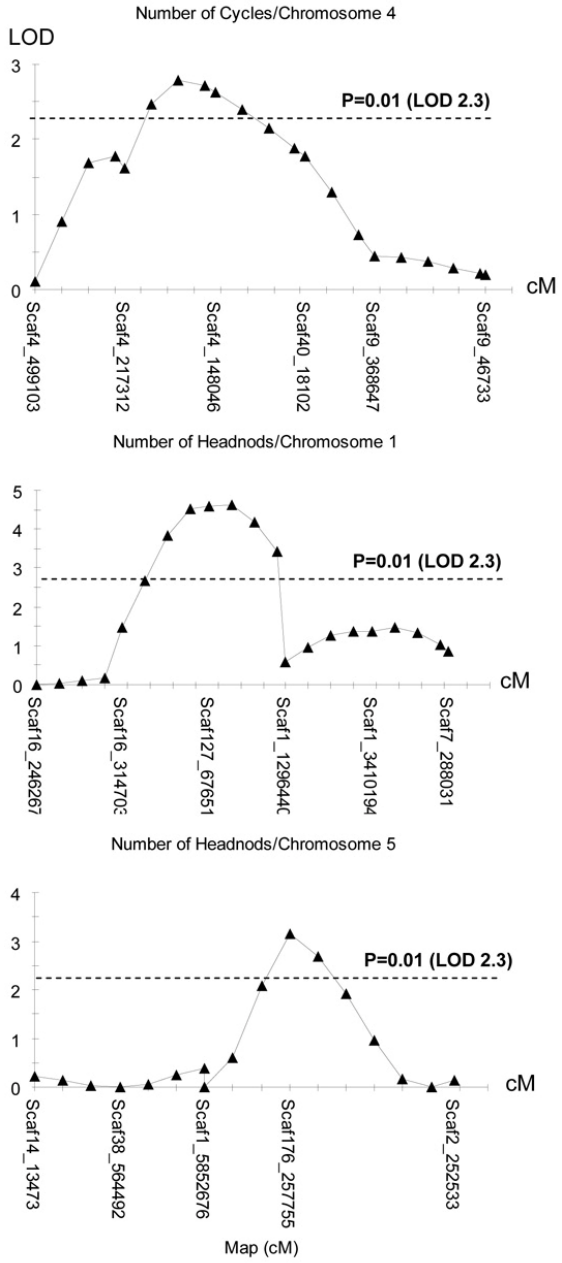


Figure S25: Three QTL for male courtship behavior. The x axis represents a chromosome (tick marks are 5cM apart) and the y axis the LOD score at 5 cM intervals using the interval mapping algorithm of MapQTL 4.0. Markers represent Microsatellite markers are at approximately equal distance (20 cM) spanning the maximum width of each chromosome and are named according to their position on the published linkage map (Fig. 2). Shown are three of the eight QTL analyses (total number of headnods, number of headnods, Table S57). The dashed line shows the genome wide significance threshold for each trait.

Supplementary Tables

Table S1: Orthologous groups (Og) present in Human and *Nasonia*, but not *Drosophila.* Available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S2: Repeats in the *Nasonia* genome assembly v1.0.
Table S3: Distribution of ANK proteins across 17 arthropods.
Table S4: List of *Nasonia* genes involved in small RNAs pathways.
Table S5: Computationally predicted miRNAs. Available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S6: Compilation of five *N. vitripennis* microRNA prediction sets. Available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S7: Yellow-like/royal jelly protein gene expression from Tiling arrays.
Table S8: Sex peptide receptors in arthropods.
Table S9: Neurohormones and their receptors in *Nasonia*.
Table S10: TRP and DEG/ENaC channel genes in the *Nasonia* genome.
Table S11: Comparision of *Apis* (Amel) and *Nasonia* (Nvit) cuticular proteins in the CPR family.
Table S12: Predicted proteins for which mass spectrometric evidence was found in larval diapause samples of *N. vitripennis*. Available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html
Table S13: Rapidly evolving genes between *Nasonia* species pairs.
Table S14: Levels of intraspecific variation in three species of *Nasonia*.
Table S15: PRANC gene expression from Tiling arrays. Number represent average expression above background threshold level.
Table S16: Venom proteins discovered by a bioinformatic and/or a proteomic approach.
Table S17: Odorant binding proteins (OBPs) in the *Nasonia* genome. Available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S18: Sequence read statistics for the *N.vitripennis* genome assembly
Table S19: Summary of gene models produced by different gene annotation pipelines.
Table S20: Comparison of GLEAN and input datasets with “gold standard” sequences.
Table S21: Comparison of GLEAN and input datasets with splice sites in EST alignments.
Table S22: *Nasonia* EST assembly summary. *Nasonia* EST assembly summary.
Table S23: EST validation of *Nasonia* gene models, with sensitivity and specificity.

Table S24: Potential genes missed by gene prediction based on orthology analysis. Available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S25: Fasta, GFF, and readme files for gene models corrected by orthology analysis. Available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S26: Validation of OGS v1.2 genes from tiling array data, from EST sequencing data and/or homology to other proteomes in the NCBI non-redundant database ($p < 1 \times 10\text{-}30$). Available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S27: Expressed mature miRNA sequences identified in the small RNA library. Available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S28: Primers used to map visible markers.
Table S29: Annotation of PILER-DF Predictions: Complete list of the 1195 PILER-DF predictions for the *N. vitripennis* genome. PILER-DF ID is indicated (column A), as well as TE Class (column B), and nearest related TE Family (column C). TE characteristics including percent tandem repeat content (colmn D), length (column E), terminal repeat type (column G), and DNA sequence (column H) are shown. alse positives were identified by BLAST alignment to known genes from *A. mellifera* and *D. melanogaster*. A FASTA of PILER-DF and Repbase sequences used for this study are shown in Table S7. Available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S30: FASTA of PILER-DF, GPS Retroid predictions & RepBase libraries used for genome masking and screening of OGS genes for contaminating repeats. Available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S31: Repeat Coordinates for Genome: Parsed output from RepeatMasker 3.25 using both PILER-DF and RepBase libraries. Scaffold start and stop coordinates are shown in column B and C. Column D indicates the Class of repeats identified. 'Novel' indicated a PILER-DF prediction with no known homology to any repeat, while 'Novel tandem' indicates a similar prediction that also has a tandem repeat content greater than 15%. Column E indicates the family of simple or interspersed repeat. The complete RepeatMasker 'out' file is also included. Available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S32: FASTA of Masked Genome Sequence using RepBase and PILER-DF libraries. Available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S33: Identified PSR repeats in *Nasonia* genome assembly v1.0.
Table S34: OGS Genes with Homology to Repeats: RefSeq, GNOMON, and GLEANR IDs from the Official Gene Set (OGS) are shown in column A. Classification of each OGS gene is shown in column B, where 'gene' indicates that no significant homology to a known repeat as found. Percent identity to PILER-DF and RepBase repeats is shown on columns D-H. False positives indicate PILER-DF predictions that matched known genes. Available online at

http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S35: Bisulfite sequencing primers for *N. vitripennis.*
Table S36: Global protein domain properties.
Table S37: List of KEGG modules including *Nasonia* lost genes.
Table S38: Gene losses in selected KEGG pathways.
Table S39: Orthologous groups (OG) present in human and *Nasonia* but not other insects.
Table S40: Runaway duplication in *Nasonia* of otherwise single copy genes.
Table S41: Losses of otherwise single copy genes in *Nasonia* (TBLASTN hit of bee protein to *Nasonia* genome not better than 10^3).
Table S42: RT-PCR and 5'- and 3'-RACE-PCR primers used to study the transformer (*tra*) gene in *N. vitripennis*.
Table S43: Identified repeat families in the *Nasonia* genome. Available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S44: Coordinates of the annotated yellow/royal jelly genes in *Nasonia*.
Table S45: Retroid elements were annotated with the Genome Parsing Suite pipeline. This is the non-redundant list of all annotated retroid elements. A FASTA of these sequences is provided in Table7. Available online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S46: List of epigenetic control related genes with different copy numbers between 4 insect species.
Table S47: Gene structure statistics for *Nasonia* in comparison to other eukaryotes
Table S48: Arthropod proteome sources for gene clustering.
Table S49: Arthropod gene model count, partitioned by orthology and paralogy.
Table S50: Hymenoptera-specific orthologs. Available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
Table S51: Overabundant *Nasonia* gene families, compared to insects.
Table S52: Evolutionary rates for *Nasonia* gene classes.
Table S53: Distribution of dN/dS across gene sizes for *Nasonia*-specific, Hymenoptera-specific, and non-specific genes.
Table S54: Genomic organization of detoxification genes.
Table S55: Scaffolds of bacterial origin: identification of a *Nasonia* commensal. Available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html
Table S56: Female and male expression variation in response to *Wolbachia* infection.
Table S57: QTL for male courtship behavior.
Table S58: Dnmt genes from *Apis*, *Ixodes*, and *Pediculus*. Available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html"

Table S2: Identified repeat classes in the *Nasonia* genome. Repeat classes are based on the Repbase classification system using RepeatMasker 3.25. Total number of repeats, total length in assembly, and percentages are shown. 'Novel' indicates a PILER-DF prediction with no known homology to any repeat, while 'Novel tandem' indicates a similar prediction that also has a tandem repeat content greater than 15%. LTR/LINE elements were classified by RT or other domains that could not distinguish between these classes.

	Number of repeats in assembly	Percent of total repeats	Total length of repeat in assembly	Percent of total repeat length
Non-Interspersed Repeats	168,379	51.9	21,360,613	43.4
Simple_repeat	96,779	29.8	4,933,404	10.0
Low_complexity	55,842	17.2	4,206,126	8.5
Satellite	10,696	3.3	7,892,239	16.0
Novel Tandem	5,062	1.6	4,328,844	8.8
Interspersed Repeats	156,315	48.1	11,547,591	56.6
Retroelements	38,765	11.9	10,425,021	21.2
LTR (retrotransposon)	18,783	5.8	7,260,751	14.8
LINE (retroposon)	18,449	5.7	2,745,796	5.6
LTR/LINE (ambiguous)	1,533	0.5	418,474	0.9
DNA Transposon	6,255	1.9	1,122,570	2.3
Novel Repeats	116,357	35.8	16,311,505	33.1

Table S3: Distribution of ANK proteins across 17 arthropods. ANK protein domains were detected in 17 arthropods using the Pfam definition (PF00023) and an E-value cutoff of 10^{-5}. ANK proteins are proteins that have at least one occurrence of an ANK domain. The ankyrin repeat is a widely spread protein-protein interaction motif, and is found in an unusually high copy number in the proteome of *N. vitripennis* (827 total instances in 205 ANK proteins). The next closest species, the honeybee *A. mellifera*, has less than half of the total number of ANK proteins found in *N. vitripennis*.

Proteome	Total proteins	Total instances
Apis mellifera	94	315
Anopheles gambiae	75	283
Aedes aegypti	92	364
Drosophila ananassae	71	264
Drosophila erecta	68	255
Drosophila grimshawi	70	256
Drosophila melanogaster	145	676
Drosophila mojavensis	69	239
Drosophila persimilis	68	204
Drosophila pseudoobscura	69	253
Drosophila sechellia	67	221
Drosophila simulans	58	191
Drosophila virilis	72	261
Drosophila willistoni	71	256
Drosophila yakuba	66	248
Nasonia vitripennis	207	827
Tribolium castaneum	124	476

Table S4: List of *Nasonia* genes involved in small RNAs pathways.

Gene	Refseq
Dicer-1	ref|XP_001605287.1
Dicer-2	ref|XP_001602524.1
Argonaute-1	ref|XP_001601049.1
Argonaute-2	ref|XP_001607164.1
	ref|XP_001607156.1
Argonaute-3	ref|XP_001603582.1
Piwi	NF
Aubergine	ref|XP_001602384.1
	ref|XP_001607362.1
	ref|XP_001605719.1
Spindle-E	ref|XP_001600067.1
Rm62	ref|XP_001604593.1
	ref|XP_001602045.1
	ref|XP_001605403.1
	ref|XP_001605420.1
r2d2	NF
Fmr1	ref|XP_001604888.1
vig	ref|XP_001602794.1
Tudor-SN	ref|NP_001153329.1
Gawky	lcl|hmm225054
Armitage	ref|XP_001605981.1
Belle	ref|XP_001605842.1
CG10883-PA	NF
CG17265-PA	ref|XP_001605643.1|
Dcp-1	ref|XP_001601191.1
Dcp-2	ref|XP_001607167.1
Drosha	ref|XP_001601748.1
Headcase	ref|XP_001607251.1
Loquacious	ref|XP_001601132.1
Pasha	lcl|hmm328774*
Pacman	ref|XP_001603129.1.
	ref|XP_001606765.1
SID-1	ref|XP_001605484.1
SID-2	NF
eri-1	ref|XP_001603726.1

Table S7: Yellow-like protein gene expression from Tiling arrays (see Materials & Methods). Numbers represent normalized average expression above background threshold level. Female R Tract = female reproductive tract, including ovaries and venom gland.

Gene Name	Gene ID (OGS v1.2)	Early embryo	Late embryo	Larvae	Pupae	Female R Tract	Testes
NvMRJPL8	XM_001599040.1	1.395	2.7925	1.955	3.3225	1.32	1.59
NvMRJPL7	XM_001599055.1	0.205	3.955	3.565	3.1925	3.905	0.51
NvMRJPL6	XM_001599068.1	0.105	1.255	0.09	0.535	2.69	0.09
Nv-yellow	XM_001599643.1	0.21	0.645	0.11	3.1025	0.475	0.25
Nv-yellow-x1c	XM_001599647.1	0.25	2.015	1.1875	2.9075	2.14	1.66
Nv-yellow-f	XM_001600594.1	1.17	4.4725	3.84	5.0725	1.21	4.84
NvMRJPL10	XM_001600620.1	0	0.32	0.465	1.235	1.69	0.775
Nv-yellow-x2	XM_001600972.1	0.345	1.37	0.8	5.415	0.455	2.63
NvMRJPL9	XM_001601334.1	0.22	0.4025	0.57	0.135	4.78	0.18
Nv-yellow-x1a	XM_001601721.1	0.1275	0.925	0.5675	1.465	2.845	1.17
Nv-yellow-x1b	XM_001601752.1	0.1575	0.8875	0.235	0.4	3.85	2.63
Nv-yellow-h	XR_036776.1	0.595	4.6775	3.81	2.935	0.74	3.16
NvMRJPL5	XM_001603256.1	0.17	1.64	1.1	3.0325	2.64	1.33
NvMRJPL4	XM_001603283.1	0.755	1.0275	0.8825	0.53	1.795	1.48
NvMRJPL3	XM_001603319.1	3.765	3.3	4.0425	2.6875	3.8	1.85
NvMRJPL1	XM_001603354.1	0.325	0.3925	0.4675	0.175	2.49	0.29
Nv-yellow-e3	XM_001603383.1	1.625	2.945	2.915	2.975	1.09	2.84
Nv-yellow-e	XM_001603435.1	0.1075	4.81	3.5375	5.1	0.18	5.205
Nv-yellow-g2b	XM_001603462.1	0.27	0.1875	0.1525	0.2225	3.83	0.66
Nv-yellow-g2a	XM_001603494.1	0.385	0.085	0.41	0.15	1.2	0.29
Nv-yellow-g2c	XM_001603523.1	0.725	0.06	0.155	0.43	3.16	0.4
Nv-yellow-g	XM_001603551.1	0	0.28	0	0.1575	1.98	0.325
Nv-yellow-b	XM_001607590.1	0.7175	3.695	2.415	4.795	1.51	3.69
Nv-yellow-x1e	XM_001607202.1	0.4125	1.3875	3.735	1.135	0.88	0.38
Nv-yellow-x1d	XM_001607460.1	0.74	3.9	2.875	2.95	1.08	1.56

Table S8: Sex peptide receptors in arthropods. The sex peptide receptor is regarded to be present if more than 50% of the amino acids in the identified sequence are identical with the *D. melanogaster* sex peptide receptor (gene no. CG16752; accession no. NP_572225.1). For *Nasonia* and *Apis*, the structurally closest receptors are myosuppressin receptors (29% identity to CG16752).

Class	Species	Present	Accession no.
Insecta (holometabola)	*Drosophila melanogaster*	+	NP_572225.1
	Drosophila secchellia	+	XP_002036913.1
	Drosophila simulans	+	XP_002106189.1
	Drosophila erecta	+	XP_001976968.1
	Drosophila yakuba	+	XP_002100008.1
	Drosophila ananassae	+	XP_001963770.1
	Drosophila persimilis	+	XP_002022959.1
	Drosophila grimshawi	+	XP_001992451.1
	Drosophila pseudoobscura	+	ABW86943.1
	Drosophila virilis	+	XP_002055695.1
	Drosophila mojavensis	+	XP_002010182.1
	Drosophila willistoni	+	XP_002071716.1
	Aedes aegypti	+	ABW86944.1
	Culex pipiens	+	XP_001849927.1
	Anopheles gambiae	+	ABW86945.1
	Bombyx mori	+	ABW86946.1
	Tribolium castaneum	+	ABW86947.1
	Nasonia vitripennis	-	
	Apis mellifera	-	
Insecta (hemimetabola)	*Acyrthosiphon pisum*	+	XP_001944453.1
	Pediculus humanus	+	XP_002424116.1
Crustacea	*Daphnia pulex*	+	gw1.17.23.1
Chelicerata	*Ixodes scapularis*	+	XP_002400964.1

Table S9: Neurohormones and their receptors in *Nasonia*. Abbreviations: AKH, adipokinetic hormone; CCAP, crustacean cardioactive peptide; DH, diuretic hormone; EH, eclosion hormone; ETH, ecdysis triggering hormone; ILP, insulin like peptide; ITP, ion transport peptide; NPF, neuropeptide F; PDF, pigment dispersing factor; PTTH, prothoracicotropic hormone; sNPF, short neuro peptide F.

Neuropeptide or protein hormone	***Nasonia* precursor gene ID**	***Nasonia* receptor gene ID**	***Drosophila* receptor orthologue (endogenous ligand)**
AKH-1	NV_13000	NV_06888	CG11325 (AKH)
AKH-2	NV_30201	NV_04132	CG11325 (AKH)
Allatostatin A	nv_13011	NV_04625	CG2872 (Allatostatin A)
Allatostatin C	NV_30011	NV_06655	CG7285 (Allatostatin C)
Allatostatin CC	NV_30012		
Bursicon alpha	nv_13010	NV_05975	CG8930 (Bursicon)
Bursicon beta	NV_03820		
CCAP	NV_04678	NV_08754	CG6111 (CCAP)
Corazonin	NV_09278	NV_10005	CG10698 (Corazonin)
DH31	NV_02135	NV_04846	CG32843 (DH31)
DH44	NV_13001	NV_07600	CG8422 (DH44)
EH	NV_08338		
ETH	NV_01589	NV_01589	CG5911 (ETH)
ILP-1	NV_03688		
ILP-2	NV_30146		
Inotocin	NV_08469	NV_02145	
ITG-like	NV_06225		
ITP	NV_07921		
Myosuppressin	NV_09202	NV_10363	CG8985 (Myosuppressin)
Neuroparsin	NV_03041		
NPF	NV_13002		
NVP-like	NV_30009		
Orkokinin	NV_30010		
PDF	Nv_15000	nv_13009	CG13758 (PDF)
PTTH	NV_30191		
Pyrokinin	NV_30144	NV_ 05180	CG8784 (Pyrokinin-2)
SIFamide	NV_30008	NV_01550	CG10823 (SIFamide)
sNPF	NV_13003	NV_06082	CG7395 (sNPF)
Tachykinin	NV_03478	NV_04616	CG7887 (Tachykinin)
		NV_09808	CG30106 (Allatostatin B)
		NV_09810	CG30106 (Allatostatin B)
		NV_08911	CG2114 (FMRFamide)
		NV_10270	CG9918 (Pyrokinin-1)
		NV_30002	CG9918 (Pyrokinin-1)
		NV_02227	CG9918 (Pyrokinin-1)
		NV_07617	CG5046
		NV_06940	CG13299
		NV_04652	CG5936
		NV_30000	CG16726
		NV_30001	CG5811
		NV_02112	CG12290
		NV_04441	CG4322
		NV_04442	CG4313
		NV_03530	CG3171

	NV_30222	CG12610
	NV_07510	
	NV_08148	
	NV_02584	
	NV_13012	
	NV_13013	
	NV_13014	
	NV_10597	CG4356 (acetylcholine)
	NV_04025	CG17004 (dopamine)
	NV_06317	CG9652 (dopamine)
	NV_10010	CG18741 (dopamine)
	NV_09647	CG18314 (dopamine, ecdysteroids)
	NV_30004	CG3856 (octopamine)
	NV_06330	CG6989 (octopamine)
	NV_03742	CG15113 (serotonin)
	NV_04419	CG12073 (serotonin)
	NV_30005	CG16766 (tyramine)
	NV_30006	CG7485 (tyramine)
	NV_06434	CG7994/CG8007
	NV_09882	CG7918
	NV_10517	CG18208
	NV_10577	CG13579

Table S10: TRP and DEG/ENaC channel genes in the *Nasonia* genome. In addition, accession numbers of the most closely related sequences (protein-protein blast) and the putative orthologs from *Drosophila melanogaster* are indicated. With one exception, the most closely related sequences are either from the honeybee *Apis mellifera* or from the beetle *Tribolium castaneum*. The function of the orthologs from *Drosophila* is also indicated.

Subfamily	Acc. number	Acc. number of most closely related sequence in GenBank	Acc. number of *Drosophila* ortholog	Synonym of *Drosophila* ortholog	Function in *Drosophila*
TRPC	XP_001605329.1	XP_001120503 (*Apis*)	CG7875	*trp*	phototransduction
	XP_001604491.1	XP_968598 (*Tribolium*)	CG18345	*trp-like*	phototransduction
	XP_001604587.1	XP_394299 (*Apis*)	CG5996	*trp gamma*	phototransduction
TRPV	XP_001606125.1	XP_625170 (*Apis*)	CG5842	*nanchung*	hearing/hygrosensation
	XP_001602588.1	XP_001121881 (*Apis*)	CG4536	*inactive*	hearing
TRPM	XP_001600197.1	XP_395829 (*Apis*)	CG30079	*trpm*	unknwown
TRPN	XP_001605939.1	XP_392309 (*Apis*)	CG11020	*nompC*	light touch/hearing
TRPA	XP_001601841.1	XP_974641 (*Tribolium*)	CG15860	*painless*	nociception
	XP_001600001.1	XP_972539 (*Tribolium*)	CG17142	*pyrexia*	geotaxis
	XP_001604029.1	XP_395234 (*Apis*)	CG31284	*water witch*	hygrosensation
	XP_001604057.1	XP_395235.2 (*Apis*)	---		
TRPML	XP_001606145.1	XP_624283 (*Apis*)	CG8743	*trpml*	TRPML
DEG/ENaC	XP_001605878	XP_001120664 (*Apis*)	CG34059	*pickpocket 16*	DEG/ENaC
	XP_001599205	XP_002055864 (*Drosophila virilis*)	CG12048	*pickpocket 21*	
	XP_001603759	XP_966654 (*Tribolium*)	CG8178	*pickpocket 4 (long form)*	
	XP_001600979	XP_001121219 (*Apis*)	CG33508	*pickpocket 13*	
	XP_001607176	XP_001122513 (*Apis*)	CG4805	*pickpocket 28 (isoform b)*	

Table S11: Comparision of *Apis* (Amel) and *Nasonia* (Nvit) cuticular proteins in the CPR family.

AmelCPR	Ortholog or paralog	NvitCPR	Percent identity in R&R consensus region(s)	Percent identity in entire protein	Match outside consensus region
1	P	1, 59			yes
2	P	1, 59			yes
3	P	9, 10			no
4	O	11	71	44	no
5	O	48	79	53	yes
6	P	several			no
7	O	50	82	53	yes
8	O	23	93	61	yes
9	O	45	82	67	yes
10	O	44	78	47	yes
11	P	45			yes
12	P	30-33			yes
13	P	30-33			yes
14	O	19	69	69	yes
15	O	25	92	58	scattered bits
16	O	42	73, 86	51	yes
17	O	18	83	55	yes
18	O	17	75	46	yes
19	O	38	91	62	yes
20	O	62	88	66	yes
21	P	21			yes
22	O	16	75	64	yes
23	O	24	91	61	yes
24	O	22	71	45	yes
25	O	52	74	61	yes
26	O	35	73	55	yes
27	O	37	92	70	yes
28	O	34	71	64	yes
29	O	53	88	67	yes
30	O	36	89	55	yes
31 frag	O	58	92		yes
32	O	61 frag	76		yes
Mean			81	58	

Table S13: Rapidly evolving gene ontology (GO) categories between *Nasonia* species pairs (Nv = *N. vitripennis*, Ng = *N. giraulti*, Nl = *N. longicornis*). In order to test for accerated evolutionwe randomly resampled dN/dS onto the gene pairs and calculated mean dN/dS values for each GO term 10,000 times to generate a null distribution for each GO term. We then compared the mean dN/dS of each GO term which appeared in at least 5 gene pairs to the null distribution to calculate a p-value. We corrected for multiple comparisons using a 5% false discovery rate with Q-value (*S82*). All GO catergories significant at the 0.01 level in at least one species pair are shown.

GO term	p-value		
	Nv/Ng	Nv/Nl	Ng/Nl
Component			
Mitochondrial large ribosomal subunit	<0.001*	<0.001*	0.002
Mitochondrial small ribosomal subunit	<0.001*	<0.001*	<0.001*
Mitochondrial respiratory chain complex I	0.028	0.004	0.962
Mitochondrial proton-transporting ATP synthase complex, coupling factor	0.031	0.036	<0.001*
Process			
RNA processing	0.001	0.005	0.483
Translation	0.003	0.029	0.001
tRNA modification	0.014	0.005	0.024
Mitochondrial electron transport, NADH to ubiquinone	0.059	0.007	0.381
ATP synthesis coupled proton transport			0.002
Function			
Structural constituent of ribosome	<0.001*	0.002	<0.001*
RNA binding	<0.001*	0.002	0.225
NADH dehydrogenase activity	0.047	0.007	0.977
NADH dehydrogenase (ubiquinone) activity	0.054	0.002	0.934
Nuclease activity		0.008	0.326

* significant at a 5% false discovery rate; blank = not enough alignments present for analysis.

Table S14: Levels of intraspecific variation in three species of *Nasonia*.

Gene/Region (Gene id)	Coding (bp)	Noncoding (bp)	species (no. of strains)	Synonymous sites		Intronic sites		Non-synonymous sites	
				π	θ	π	θ	π	θ
similar to ENSANGP00000013532 (LOC100119230)	765	0	NV(19)	0.0035	0.0033	NA	NA	0.0000	0.0000
			NL(16)	0.0022	0.0052	NA	NA	0.0000	0.0000
			NG(13)	0.0000	0.0000	NA	NA	0.0000	0.0000
similar to ENSANGP00000031746 (LOC100119303)	462	253	NV(19)	0.0000	0.0000	0.0010	0.0010	0.0000	0.0000
			NL(16)	0.0013	0.0030	0.0000	0.0000	0.0008	0.0017
			NG(13)	0.0003	0.0000	0.0006	0.0013	0.0000	0.0000
similar to cytochrome P450 CYP4AB1 (LOC100123049)	414	375	NV(19)	0.0068	0.0063	0.0015	0.0031	0.0014	0.0024
			NL(12)	0.0039	0.0067	0.0003	0.0010	0.0014	0.0008
			NG(8)	0.0019	0.0037	0.0000	0.0000	0.0009	0.0018
similar to alpha-glucosidase isozyme I (LOC100121102)	345	307	NV(6)	0.0000	0.0000	0.0020	0.0010	0.0000	0.0000
			NL(6)	0.0000	0.0000	0.0010	0.0010	0.0000	0.0000
			NG(7)	0.0020	0.0020	0.0070	0.0060	0.0000	0.0000
similar to p27BBP/eIF6-like (LOC100115084)	261	367	NV(19)	0.0082	0.0045	0.0006	0.0016	0.0000	0.0000
			NL(16)	0.0000	0.0000	0.0037	0.0025	0.0000	0.0000
			NG(8)	0.0000	0.0000	0.0004	0.0008	0.0000	0.0000
similar to ENSANGP00000029450/cg31 (LOC100116515)	339	198	NV(10)	0.0000	0.0000	0.0010	0.0020	0.0000	0.0000
			NL(9)	0.0000	0.0000	0.0010	0.0020	0.0000	0.0000
			NG(8)	0.0000	0.0000	0.0010	0.0010	0.0000	0.0000
similar to NADH dehydrogenase 2C putative (LOC100115994)	132	396	NV(9)	0.0000	0.0000	0.0030	0.0050	0.0050	0.0030
			NL(10)	0.0000	0.0000	0.0010	0.0020	0.0000	0.0000
			NG(7)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
similar to Phospholipase A-2-activating protein (LOC100122485)	699	0	NV(8)	0.0040	0.0050	NA	NA	0.0000	0.0000
			NL(9)	0.0000	0.0000	NA	NA	0.0000	0.0000
			NG(5)	0.0000	0.0000	NA	NA	0.0000	0.0000
similar to ENSANGP00000017418/replication factor (LOC100114948)	303	294	NV(5)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
			NL(7)	0.0000	0.0000	0.0010	0.0020	0.0000	0.0000
			NG(6)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
similar to 60S ribosomal protein L7/L12	411	87	NV(19)	0.0000	0.0000	0.0022	0.0022	0.0000	0.0000

(LOC100118391)			NL(16)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
			NG(7)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
similar to Lipase	249	150	NV(18)	0.0000	0.0000	0.0000	0.0000	0.0020	0.0030
(LOC100118843)			NL(14)	0.0080	0.0060	0.0000	0.0000	0.0020	0.0020
			NG(14)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
similar to arp2/3 complex	324	377	NV(18)	0.0003	0.0010	0.0010	0.0010	0.0000	0.0000
(LOC100118949)			NL(14)	0.0000	0.0000	0.0020	0.0020	0.0000	0.0000
			NG(14)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ATP synthase,coupling factor F	321	443	NV(19)	0.0000	0.0000	0.0040	0.0030	0.0000	0.0000
(LOC100114299)			NL(16)	0.0000	0.0000	0.0020	0.0010	0.0020	0.0010
			NG(14)	0.0000	0.0000	0.0003	0.0010	0.0000	0.0000
similar to Casein Kinase I, alpha I	225	673	NV(19)	0.0040	0.0100	0.0002	0.0004	0.0010	0.0010
(LOC100114338)			NL(16)	0.0020	0.0010	0.0020	0.0010	0.0080	0.0060
			NG(14)	0.0000	0.0000	0.0000	0.0000	0.0010	0.0010
similar to fumarylacetoacetate hydrolase	450	270	NV(19)	0.0040	0.0010	0.0020	0.0020	0.0010	0.0010
(LOC100117860)			NL(16)	0.0110	0.0110	0.0010	0.0010	0.0000	0.0000
			NG(14)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
similar to long wavelength sensitive Opsin I	516	146	NV(19)	0.0010	0.0020	0.0060	0.0040	0.0010	0.0010
(LOC100122456)			NL(16)	0.0100	0.0100	0.0200	0.0400	0.0030	0.0030
			NG(14)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
similar to Phosphoglucose isomerase	281	159	NV(19)	0.0000	0.0000	0.0010	0.0030	0.0010	0.0010
(LOC100115189)			NL(16)	0.0100	0.0100	0.0100	0.0100	0.0020	0.0030
			NG(14)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
similar to CG1970-PA	780	0	NV(16)	0.0006	0.0016	NA	NA	0.0000	0.0000
(LOC100115931)			NL(15)	0.0000	0.0000	NA	NA	0.0000	0.0000
			NG(12)	0.0000	0.0000	NA	NA	0.0000	0.0000
similar to ENSANGP00000025535	353	171	NV(17)	0.0183	0.0148	0.0037	0.0035	0.0000	0.0000
LOC100121837			NL(11)	0.0023	0.0042	0.0000	0.0000	0.0000	0.0000
			NG(11)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
similar to Juvenile hormone-inducible protein	794	63	NV(19)	0.0040	0.0052	0.0000	0.0000	0.0008	0.0009
LOC100121269			NL(16)	0.0000	0.0000	0.0020	0.0048	0.0002	0.0005
			NG(13)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
similar to Replication protein A middle subunit	220	428	NV(19)	0.0038	0.0054	0.0041	0.0048	0.0000	0.0000

LOC100119335			NL(16)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
			NG(13)	0.0000	0.0000	0.0007	0.0008	0.0000	0.0000
similar to putative accessory gland protein	487	220	NV(19)	0.0009	0.0023	0.0046	0.0039	0.0017	0.0020
LOC100119650			NL(17)	0.0010	0.0026	0.0024	0.0013	0.0013	0.0014
			NG(14)	0.0039	0.0028	0.0000	0.0000	0.0000	0.0000
similar to androgen induced inhibitor	729	140	NV(10)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
of proliferation (as3) / pds5			NL(14)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
LOC100118251			NG(12)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
similar to mitochondrial processing peptidase	603	129	NV(14)	0.0021	0.0046	0.0021	0.0024	0.0000	0.0000
alpha subunit LOC100120518			NL(15)	0.0010	0.0023	0.0000	0.0000	0.0000	0.0000
			NG(13)	0.0028	0.0024	0.0022	0.0025	0.0003	0.0007
Total	10463	5646							
Average			NV	0.0026	0.0028	0.0019	0.0021	0.0006	0.0006
			NL	0.0022	0.0026	0.0024	0.0034	0.0009	0.0008
			NG	0.0005	0.0005	0.0006	0.0006	0.0001	0.0001

π = the average number of nucleotide substitutions per site between two sequences; θ = segregating sites.

Table 15: PRANC gene expression from Tiling arrays. Number represent average expression above background threshold level.

Gene ID (OGS v1.2)	Early embryo	Late embryo	Larvae	Pupae	Female	Male
XP_001604063.1	1.13	1.71	1.07	1.58	1.32	1.31
XP_001603925.1	2.51	3.86	3.37	3.58	3.31	3.27
XP_001603132.1	1.62	1.09	2.75	2.26	1.75	1.87
XP_001602219.1	0.88	1.72	2.22	1.59	1.44	1.76
XP_001601419.1	3.99	4.02	3.75	4.17	3.81	4.08
GLEAN_00513	0.00	0.16	0.21	0.09	0.13	0.10
GLEAN_00563	3.43	0.66	0.69	0.48	1.04	1.48
GLEAN_03473	2.43	3.18	3.62	3.48	3.34	3.34
GLEAN_07757	2.78	2.26	2.05	2.20	2.15	2.28
GLEAN_10854	1.53	1.26	1.27	1.58	1.35	1.30
GLEAN_13854	0.04	0.43	0.32	0.57	0.30	0.41
hmm308494	2.86	2.78	3.07	3.55	2.99	3.01
hmm708344	0.48	0.89	0.30	1.97	1.52	1.33

Table S16: Venom proteins discovered by a bioinformatic and/or a proteomic approach. Protein name, Genbank accession number, methods of discovery, and ratio of female reproductive tract: testes expression level in the Tiling array experiments, dN/dS ratio between *N. vitripennis* and *N. giraulti*, Hymenoptera species with a similar venom protein, and putative venom-specific function are listed.

Protein name	Acc. No.	Method[a]	Female/ Testes	dN/dS	Other species[b]	Putative venom-specific function[c]
Proteases and peptidases						
metalloprotease	XP_001604362	B1, B2	3.79	0.82	*Eulophus pennicornis*, *Melittobia digitata*, *Pimpla hypochondriaca*	host immune suppression
angiotensin-converting enzyme	XP_001607198	B1, B2	1.50	0.09	*Pimpla hypochondriaca*	processing of peptide precursors
dipeptidylpeptidase IV	XP_001599462	P2	2.01		*Apis mellifera*	processing of peptide precursors
serine protease	lcl|hmm408134	P2	17.89		*Apis mellifera, Cyphononyx dorsalis*, *Pimpla hypochondriaca*	n.k.
serine protease	lcl|hmm536144	P2	16.29		see above	n.k.
serine protease	XP_001604852	P2	10.39	1.93	see above	n.k.
serine protease	lcl|hmm315294	P2	25.15		see above	n.k.
serine protease	XP_001600774	P2	7.83	0.92	see above	n.k.
serine protease	XP_001600807	B1, P2	7.32	1.38	see above	n.k.
serine protease	XP_001600838	B1, P2	8.41		see above	n.k.
serine protease (SP50)	XP_001602769	B1, B2	1.32	0.98	see above	n.k.
serine protease (SP97)	XP_001604873	P1, P2	21.28		see above	n.k.
serine protease (SP101)	XP_001605133	B1, B2	2.27	0.33	see above	n.k.
serine protease/CLIP	XP_001600149	P2	20.65	1.21	see above	n.k.
serine protease/CLIP	XP_001600178	B1, P2	4.08	0.14	see above	n.k.
serine protease/CLIP	XP_001599920	B1, P2	4.96	0.09	see above	n.k.
serine protease/CUB (SPH42)	XP_001602368	B1, B2	3.92	0.17	see above	n.k.
serine protease homolog (SPH21)	XP_001600151	B1, B2, P2	2.84	0.28	*Cotesia rubecula*	host immune suppression; negative regulation of PO activation; inhibition melanisation host hemolymph
Protease inhibitors						
cysteine-rich/KU venom protein	XP_001604564	B1, B2	0.98		*Pimpla hypochondriaca*	n.k.

cysteine-rich/pacifastin venom protein 1	XP_001606763	B1, B2	4.15	0.35	*Pimpla hypochondriaca*	n.k.
cysteine-rich/pacifastin venom protein 2	XP_001606768	B1, B2	16.54	0.31	see above	n.k.
cysteine-rich/TIL venom protein 1	XP_001607367	P1, P2	27.38		*Pimpla hypochondriaca*, *Microplitis demolitor* bracovirus	stabilization or inhibition of venom phenoloxidase
cysteine-rich/TIL venom protein 2	XP_001607359	B1, B2	0.96		see above	see above
Kazal type serine protease inhibitor-like venom protein 1	XP_001604686	P2	2.21	0.26	no	n.k.
Kazal type serine protease inhibitor-like venom protein 2	lcl|hmm867024	P2	10.99		no	n.k.
small serine proteinase inhibitor-like venom protein	XP_001607610	P2	240.00		no	n.k.
Carbohydrate metabolism						
chitinase 5	XP_001602780	B1, B2	0.56	0.05	*Chelonus* sp., *Toxoneuron nigriceps*	n.k.
glucose dehydrogenase-like venom protein	XP_001600948	P2	23.61	0.32	no	n.k.
trehalase	XP_001602179	B1, B2	1.65	0.16	*Pimpla hypochondriaca*	nutritional function
DNA metabolism						
apyrase	XP_001603046	P2	1.48	0.52	no	n.k.
endonuclease-like venom protein	XP_001599850	P2	38.38	1.51	no	n.k.
inosine-uridine preferring nucleoside hydrolase	XP_001599345	P2	0.35	0.54	no	n.k.
Glutathione metabolism						
γ-glutamyl transpeptidase-like venom protein 1	XP_001607488	B1, B2, P1, P2	30.91	0.19	*Aphidius ervi*	apoptosis of host germaria and ovariole sheath cells
γ-glutamyl transpeptidase-like venom protein 2	XP_001604839	B1, B2, P2	21.00	0.48	see above	see above
γ-glutamyl cyclotransferase-like venom protein	XP_001605740	P2	0.80	0.76	no	n.k.
Esterases						
acid phosphatase	XP_001600562	B1, B2	0.55	0.39	*Apis mellifera*, *Pimpla hypochondriaca*, *Pteromalus*	n.k.

					puparum	
acid phosphatase	XP_001605498	B1, P2	6.51	0.37	see above	n.k.
multiple inositol polyphosphate phosphatase-like venom protein	XP_001605344	P2	2.34	0.41		
arylsulfatase b	XP_001603886	P2	3.82	0.36	no	n.k.
lipase	lcl|hmm642184	P1, P2	37.23		*Pimpla hypochondriaca*	n.k.
lipase-like venom protein	lcl|hmm589104	P2	14.69		see above	n.k.
α-esterase	XP_001602279	P2	3.75	0.51	see above	n.k.
Recognition/binding proteins						
β-1,3-glucan recognition protein	XP_001607754	P2	2.16	0.13	no	n.k.
chitin binding protein-like venom protein	lcl|hmm464144	P2	5.43		no	n.k.
GOBP-like venom protein	XP_001604989	P1, P2	16.87	0.72	no	n.k.
low-density lipoprotein receptor-like venom protein	lcl|hmm94654	P2	4.97		no	n.k.
Immune related proteins						
calreticulin	XP_001600192	B1, B2	1.01		*Cotesia rubecula*	host immune suppression; competes for binding sites with host hemocyte calreticulin, known to mediate early-encapsulation reactions
C1q-like venom protein	XP_001608267	B1, B2	0.66		*Apis mellifera*	n.k.
IG-like venom protein	XP_001608198	P1, P2	2.83	0.11	no	suppressor of insulin-mediated growth
Others						
aminotransferase-like venom protein 1	XP_001607226	P1, P2	13.64		no	n.k.
aminotransferase-like venom protein 2	XP_001602041	P1, P2	0.00	0.22	no	n.k.
antigen 5-like protein 1	XP_001603611	B1, B2	0.83	0.58	*Vespula* spp., *Vespa* spp., *Dolichovespula* spp., *Solenopsis* spp.	n.k.
antigen 5-like protein 2	XP_001603715	B1, B2	1.61		see above	
aspartylglucosaminidase	XP_001603206	B1, B2	0.91	0.14	*Asobara tabida*	host paralysis
laccase	XP_001605351	B1, B2	0.91	0.29	*Pimpla hypochondriaca*	phenoloxidase activity; L-DOPA oxidising activity
laccase	XP_001600917	B1, P2	50.15	0.36	see above	see above

Unknown						
venom protein D	lcl|hmm895104	P2	1.82		no	n.k.
venom protein E	XP_001603438	P2	11.45		no	n.k.
venom protein F	XP_001602939	P2	0.94		no	n.k.
venom protein G	lcl|hmm716344	P2	7.89		no	n.k.
venom protein H	lcl|hmm214614	P2	47.00		no	n.k.
venom protein I	lcl|hmm282804	P2	28.87		no	n.k.
venom protein J	lcl|hmm34174	P2	42.00		no	n.k.
venom protein K	lcl|hmm320864	P2	17.24		no	n.k.
venom protein L	lcl|hmm726814	P2	9.44		no	n.k.
venom protein M	lcl|hmm72024	P2	6.58		no	n.k.
venom protein N	lcl|hmm203054	P2	2.28		no	n.k.
venom protein O	lcl|hmm169294	P2	0.71		no	n.k.
venom protein P	XP_001606165	P2	2.52	0.40	no	n.k.
venom protein Q	XP_001607549	P2	6.47	0.29	no	n.k.
venom protein R	XP_001603774	P2	2.28	0.66	no	n.k.
venom protein S	XP_001606832	P2	2.43	0.34	no	n.k.
venom protein T	XP_001601803	P2	0.49	0.71	no	n.k.
venom protein U	lcl|hmm215044	P1, P2	83.00		no	n.k.
venom protein V	lcl|hmm734344	P1, P2	23.27		no	n.k.
venom protein W	XP_001602724	P1, P2	2.60		no	n.k.
venom protein X	XP_001605793	P1, P2	2.48	1.81	no	n.k.
venom protein Y	XP_001603600	P1, P2	12.98	0.86	no	n.k.
venom protein Z	XP_001607601	P1, P2	14.24	2.03	no	n.k.

[a] B, bioinformatic approach; B1, BLAST-search; B2, subsequent RT-PCR; P, proteomic approach; P1, 2D-LC-MALDI TOF MS; P2, 2D-LC-ESI-FT-ICR MS

[b] Hymenoptera species with similar venom protein. In case of the serine protease group we only differentiated between the active enzymes (with or without additional CLIP/CUB domain) and the serine protease homolog. We did not distinguished between lipase, lipase-like venom protein and α-esterase

[c] n.k., not known

Table S18: Sequence read statistics for the *N.vitripennis* genome assembly

Insert Size	Raw Reads	Passed Reads	Passed Sequence (Mbp)	% Mate Pair Success	Clone Coverage	Assembled Reads	Assembled Coverage
3-6kb	3,189,565	2,878,480	2,110.7	96.6	26X	2,011,187	5.6X
30-40kb	172,140	140,873	98.5	96.4	10X	101,853	0.32X
130-140kb	104,722	89,201	59.4	99.6	24X	69,168	0.2X
Total	3,466,427	3,108,554	2,268.7	96.7	60X	2,182,208	6.12X

Table S19: Summary of gene models produced by different gene annotation pipelines. NCBI models are subdivided into protein coding models completely or partially based on EST or protein alignments, pseudogene models containing debilitating frameshift or nonsense codons, and *ab initio* models

Annotation pipeline	Number of gene models
NCBI	
complete support	1,597 (1,688 transcripts)
partial support	7,563
pseudogenes	1,575
ab initio	16,459
AUGUSTUS	30,196
FgenesH	32,405
FgenesH++	26,057 (26,115 transcripts)
GENEid	15,170
GLEAN	15,216
GLEAN (- RefSeq)	6,935
OGS v1.2	18,850 (18,941 transcripts) (these numbers include pseudogenes) 17,279 (17,370 transcripts) (without pseudogenes)

Table S20: Comparison of GLEAN and input datasets with "gold standard" sequences.

	Genes[a]	Gene Models[b]	Perfect matches to GS[cd]	High Identity matches to GS[ce]	No. RefSeq that overlap GLEAN[f]	Percent of Column I with > 1 overlapping GLEAN[g]	Input to GLEAN[h]
NCBI RefSeq	9230	9321	2	11	NA	NA	NA
NCBI ab initio	16324	16415	2	52	NA	NA	NA
Augustus	30196	30196	5	55	NA	NA	NA
Fgenesh	32417	32417	3	66	NA	NA	NA
Fgenesh++	26057	26115	4	60	NA	NA	NA
GLEAN5	22844	22844	4	51	8790	7.4	RefSeq, Augustus,Fgenesh
GLEAN6	15216	15216	2	53	8486	1.8	RefSeq, NCBI Ab initio, Augustus, Fgenesh
GLEAN7	28248	28248	3	58	8800	13.9	RefSeq, NCBI Ab initio, Augustus, Fgenesh, Fgenesh++

[a]Number of protein coding gene loci, computed after grouping gene models with overlapping CDS.
[b]Number of protein coding transcripts.
[c]GS = Gold Standard set consisting of 82 expert annotated coding sequences.
[d]Perfect match was defined as ≥99% identity over the entire length of both sequences.
[e]High identity match was defined as 99% identity with no length limitation.
[f]The number of RefSeq genes that overlap GLEAN genes were computed using genome coordinates.
[g]The proportion of RefSeq genes overlapping GLEAN that overlap with more than one GLEAN gene. This suggests that the larger number of gene models in GLEAN5 and GLEAN7 compared to GLEAN6 was due to splitting genes in GLEAN5 and GLEAN7.
[h]In addition to gene prediction sets, alignments of Nasonia ESTs and SwissProt metazoan protein homologs were used in all GLEAN analyses.

Table S21: Comparison of GLEAN and input datasets with splice sites in EST alignments.

	Unique predicted donor/acceptor sites	Internal EST donor/acceptor sites[a]	Perfect matches to EST donor/acceptor site	Perfect matches per internal EST donor/acceptor site	Perfect matches per predicted donor/acceptor site	Donor matches	Acceptor matches
NCBI RefSEq	52820	7109	6270	0.8819	0.1187	6315	6319
NCBI ab initio[b]	29930	83	67	0.8072	0.0022	78	73
Augustus	82987	7412	6555	0.8844	0.0790	6663	6662
Fgenesh	97881	7181	5971	0.8315	0.0610	6341	6288
Fgenesh++	89940	7085	5913	0.8345	0.0657	6207	6181
GLEAN5	72481	7121	6353	0.8921	0.0877	6444	6440
GLEAN6	75398	7141	6349	0.8891	0.0842	6447	6411
GLEAN7	91039	7000	6193	0.8847	0.0680	6383	6344

[a]Internal cDNA donor/acceptor sites are the number of cDNA splice junctions that align between the start and stop codons of gene models. The total number of cDNA splice junctions is 7,698.

[b]Note that the NCBI ab initio set is in total disagreement with the homology supported RefSeq set. They are each non-overlapping subsets resulting from the NCBI Gnomon pipeline. On the other hand, Augustus, Fgenesh, and Fgenesh++ are more similar to the combined Gnomon set, and thus more likely than NCBI ab initio to include gene models with homology to known genes. This explains the seemingly poor performance of NCBI ab initio compared to the other input gene sets.

Table S22: *Nasonia* EST assembly summary. *Nasonia* EST assembly summary. *Nasonia* ESTs were collected from GenBank on 2008/11/14, and assembled with PASA. These and gene model validations are available at http://insects.eugenes.org/arthropods/data/nasonia/pasa_est/.

Nasonia EST Assemblies	Count
Total EST sequences	175853
Nasonia vitripennis	145793
Nasonia giraulti	30060
ESTs with any alignment	167823
Valid EST alignments	147382
Assemblies	21865
Clusters (distinct genes)	17847

Table S23: EST validation of *Nasonia* gene models, with sensitivity and specificity. EST data as in Table S2.

EST Status	OGS v1.2	Glean6	Gnomon	Fgenesh	Augustus
Incorporated	5707	1297	5750	2057	1494
UTR added	812	5028	706	4732	5569
Gene extension	359	324	372	552	405
Int. structure	2781	2728	2809	2413	2503
Gene merge	707	432	824	935	855
Alternate splice	2036	1990	2037	2155	2161
Genes with EST	10366	9809	10461	10689	10826
Total Genes	18941	15216	27431	26115	30196
Discrepancy	0.371	0.355	0.383	0.365	0.348
Specificity	0.547	0.645	0.381	0.409	0.359
Sensitivity	0.717	0.679	0.724	0.740	0.749

Status Key: Inc = Incorporated, UTR add = UTR addition, Gene ext = Gene extension; Int structure = Internal gene structure rearrangement, Gene merge = Gene merging, Alt. splice = Alternate splicing isoform. Categories of Gene extension, Int. structure chage and Gene merge represent potential gene model mistakes. Discrepancy is calculated as the sum of clusters supporting a gene model mistake /Total Genes with ESTs. Specificity (Sp) is calculated as EST found/Total Genes, Sensitivity (Sp) is calculated as EST found/Total EST clusters.

Table S28: Primers used to map visible markers. All amplify length-polymorphism regions (indels) between *N. giraulti* and *N. vitripennis*.

Primer name	Primer sequence	Scaffold	Position
	Chromosome 5 (*R-locus*)		
S2-i5027-F1	TACGCAGCTGACCCAAATGT	2	5,027,000
S2-i5027-R2	ATGTGATGGCTATACTGCTATTTGG	2	5,027,000
S2-i7194-F1	ATTTTCTAATGAATCGCCGC	2	7,194,000
S2-i7194-R1	TGTGCACTATTGATTTTTCGATAGG	2	7,194,000
S163-i0251-F1	ATTAACTCGCCGGAGGGAC	163	251,000
S163-i0251-R2	TGTCCGTCCTCGCTTACTTCT	163	251,000
S10-i0005-F1	GAAAAATCTGCCAGCAGCC	10	5,000
S10-i0005-R1	AGTAGCGCGAGATCGACTGAG	10	5,000
S10-i1406-F2	AAGTGTCCATACCTACAACTGCATC	10	1,406,000
S10-i1406-R1	GACTCCTCAAGATGTTTCGCAG	10	1,406,000
S27-i1432-F1	GGGTAAATAAGCTCTGGATTGTGA	27	1,432,000
S27-i1432-R2	GAGGCATGGAAGCTTGAATC	27	1,432,000
	Chromosome 3 (*st318, mm*)		
S17-i1078-F1	ATTGGGTTGAATACTACTTTTGTTGT	17	1,078,000
S17-i1078-R1	CTTACAGAGTGGTGGGGGTAGAG	17	1,078,000
S17-i1504-F1	AGAGTATTGATTGAAAAACGTGTGC	17	1,504,000
S17-i1504-R1	CAGGCATTCCAAGCGAAGAG	17	1,504,000
S22-i1033-F1	CATTGAAGAAGCTGACTCTCGG	22	1,033,000
S22-i1033-R1	GAATTTCACTGCTTCGAACTATACAG	22	1,033,000
S22-i2810-F2	CACGAATCGAATAAGATATGGGAG	22	2,810,000
S22-i2810-R2	GTCTTCATTCGAAATGATCTGTACTT	22	2,810,000

PCR conditions for all assays: 94° C for 2 min.; 36 cycles of 94° C for 30 sec., 55° C for 60 sec., 72° C for 60 sec.; 72° C for 10min.

Table S33: Identified PSR repeats in *Nasonia* genome assembly v1.0. Published repeats were downloaded from NCBI and compared to the *Nasonia vitripennis* V1 genome assembly using WU-BLASTN. In some cases the complete, canonical repeat could not be detected in the assembly.

	Number of elements	Total length of elements (bp)	Number of scaffolds with repeat	Range of canonical repeat lengths found in scaffolds	Length of canonical repeat (bp)
AAAGTCT[T/C]GACTT	28	364	29	100%	13
psr105-3	96	2553	290	9-17%	214
psr18-1	114	3021	82	8-17%	213
psr22-2	168	4200	113	8-24%	183
psr2-1	340	7938	142	9-23%	171
nv79-16a	2973	61209	500	14-91%	94
nv85-7	734	94471	162	9-101%	175
nv126-6	4807	268838	588	12-105%	110
nv104-6	2520	357525	210	11-105%	162
TOTAL	11780	800119			

Table S35: Bisulfite sequencing primers for *N. vitripennis.*

Gene name	Gene ID	Exon	CpG O/E	Amplicon length (bp)	Primer ID	Primer sequence (5'-->3')	Frequency of methylated CpGs*
Vitellogenin	XM_001607338.1	3	1.00	360	Nbp1-f	TGAATTGAAGGATTTTGTATTTTTT	2/230
					Nbp1-r	TAATCCTAAACTTCCCCATAAAAA	
Epithelial membrane protein	XM_001600728.1	3	1.58	357	Nbp2-f	TTGAAGATTAAGTTTTTYGTTGAG	5/261
					Nbp2-r	CAAAACCCTACAAAAATCCTTATC	
Polypeptide of 976 aa	XM_001600593.1	4	0.46	313	Nbp3-f	GGAGGAAGAGTTATGAAAATTTAT	16/18
					Nbp3-r	CATTTTAAAACATAAACATAACATCTC	
eIF 2a kinase	XM_001606530.1	4	0.42	339	Nbp4-f	TGTTAATTTTAAATGGTTTGTAATGT	32/50
					Nbp4-r	TTTCTTCCRTACTAAACCTAAAAC	
eIF2B-gamma protein	XM_001601041.1	3	0.36	264	Nbp5-f	GAAATTTTAGTATTTTGAAAGGTGAA	38/45
					Nbp5-r	CAAATTTCCCATTCATAACATAAA	

*calculated as (the number of methylated CpGs/total number of CpGs surveyed)

Table S36: Global protein domain properties. Domain coverage refers to the fraction of proteins for which at least one domain was detected (see SOM Materials and Methods V.5 for the annotation procedure), multiple domain proteins to the fraction of domains with at least two domains, and unique domain arrangements to the number of unique domain arrangement, i.e. arrangements (including single domain proteins) which occur in the whole proteome.

Species	Total number of domains/proteins	Domain coverage of proteom	Multiple domain proteins	Unique domain arrangements
N. vitripennis	4636/19800	80%	41%	5026
A. aegypti	5062/15419	73%	34%	4934
A. gambiae	4982/12457	75%	36%	4734
A. mellifera	5087/11062	79%	40%	5454
D. melanogaster	5900/21127	79%	44%	5704

Table S37: List of KEGG modules including *Nasonia* lost genes. Lost genes in *Nasonia* are identified through reference key genomes (*Apis mellifera, Pediculus humanus, Tribolium castaneum, Drosophila melanogaster, Daphnia pulex, Homo sapiens*). Module IDs written in red represent validated ones by lost gene enrichment p-value at module-level (7th column) or pathway-level. Asterisks (*) in the table indicate modules belonging to amino acid metabolism. **Bold** modules were significantly enriched.

	Descriptions	Lost enzymes (EC) from *Nasonia*	OG for lost enzymes	Number of lost enzymes	Number of enzymes in the module	Enrichment p-value (by hypergeometric test)	Dunn-Sidak correction (background: total number of modules=275)	Dunn-Sidak correction (background:the number of modules including lost enzymes=36)
M00031*	**Glycine biosynthesis, serine => glycine**	2.1.2.1	2707	1	1	0	TRUE	TRUE
M00079*	**Histidine degradation, histidine => N-formiminoglutamate => glutamate**	3.5.2.7 4.2.1.49 4.3.1.3	1713 2160 4929	3	3	0	TRUE	TRUE
M00082*	**Selenocysteine metabolism**	4.4.1.16	5382	1	1	0	TRUE	TRUE
M00104	**CMP-N-Acetylneuraminate biosynthesis (mammals), ManNAc => Neu5Ac-9P => CMP-Neu5Ac**	2.5.1.56 2.5.1.57 2.7.7.43	5945 8764	3	3	0	TRUE	TRUE
M00162	**Eicosanoid biosynthesis, arachidonate => PGH2**	1.14.99.1	4150	1	1	0	TRUE	TRUE
M00170	**Eicosanoid biosynthesis, PGH2 => TX**	5.3.99.5	8006	1	1	0	TRUE	TRUE
M00207	**Phosphatidylserine (PS) biosynthesis, PE => PS**	2.7.8.-	3342	1	1	0	TRUE	TRUE
M00208	**Phosphatidylserine (PS) biosynthesis, PC => PS**	2.7.8.-	3342	1	1	0	TRUE	TRUE
M00211	**Diphosphatidylglycerol biosynthesis, CDP-glycerol => cardiolipin**	2.7.8.-	3342	1	1	0	TRUE	TRUE
M00222	**C30 isoprenoid biosynthesis, squalene => lanosterol**	1.14.99.7	9504	1	1	0	TRUE	TRUE
M00223	**C30 isoprenoid biosynthesis, squalene => cycloartenol**	1.14.99.7	9504	1	1	0	TRUE	TRUE
M00068*	**Tryptophan metabolism, kynurenine => 2-aminomuconate**	1.13.11.6 3.5.1.9 3.7.1.3	5433 7331 9052	3	5	0.000986424	FALSE	TRUE
M00013	Glyoxylate biosynthesis, glycolate => glyoxylate	1.1.1.26 1.1.1.79	6095	2	3	0.0018159	FALSE	FALSE
M00270	C1-unit interconversion	2.1.2.1 4.3.1.4 6.3.3.2	2707 5388 8703	3	7	0.005657907	FALSE	FALSE
M00042	Urea cycle	2.1.3.3 3.5.3.1	6795 9221	2	4	0.0066094	FALSE	FALSE
M00111	Glycosaminoglycan biosynthesis, a common tetrasaccharide	2.4.1.134 2.4.1.135	7102 7635	2	4	0.0066094	FALSE	FALSE

M00027*	Valine biosynthesis, pyruvate => valine	2.2.1.6	3214	1	2	0.015001624	FALSE	FALSE
M00028*	Leucine biosynthesis, pyruvate => leucine	2.2.1.6	3214	1	2	0.015001624	FALSE	FALSE
M00029*	Isoleucine biosynthesis, pyruvate => isoleucine	2.2.1.6	3214	1	2	0.015001624	FALSE	FALSE
M00125	Cerebroside and Sulfatide biosynthesis	2.8.2.11	10220	1	2	0.015001624	FALSE	FALSE
M00202	Acylglycerol degradation	3.1.1.23	9705	1	2	0.015001624	FALSE	FALSE
M00250	Riboflavin biosynthesis, GTP => riboflavin/FMN/FAD	2.7.7.2	5239	1	2	0.015001624	FALSE	FALSE
M00066*	Tryptophan metabolism, tryptophan => serotonin => 5-hydroxyindoleacetate/melatonin	1.4.3.4 2.3.1.87	8327 9447	2	5	0.015043865	FALSE	FALSE
M00067*	Tryptophan metabolism, tryptophan => kynurenine => kynurenate	3.5.1.9	9052	1	3	0.041373072	FALSE	FALSE
M00074	Tyrosine metabolism, tyrosine => L-DOPA => dopaquinone => melanin	5.3.3.12	6875	1	3	0.041373072	FALSE	FALSE
M00076	Dopamine / noradrenaline / adrenaline metabolism	1.4.3.4	8327	1	3	0.041373072	FALSE	FALSE
M00115	Fucose biosynthesis, GDP-mannose => L-fucose	2.7.1.52	6030	1	3	0.041373072	FALSE	FALSE
M00152	Dermatan sulfate degradation	3.2.1.76	5971	1	3	0.041373072	FALSE	FALSE
M00237	Vitamin K cycle	1.6.5.2	11956	1	3	0.041373072	FALSE	FALSE
M00262	Ptrescine metabolism, N-acetylation, putrescine => 4-aminobutanoate	1.4.3.4	8327	1	3	0.041373072	FALSE	FALSE
M00151	N-glycan biosynthesis, complex type	2.4.1.144 2.4.99.1	7136 7239	2	7	0.043723179	FALSE	FALSE
M00078	Tyrosine degradation, tyrosine => homogentisate	5.2.1.2	5385	1	4	0.076136744	FALSE	FALSE
M00224	Choresterol biosynthesis	1.3.1.21	6647	1	4	0.076136744	FALSE	FALSE
M00174	beta-Oxidation	1.3.99.13 1.3.99.2	4203 4379	2	10	0.113967097	FALSE	FALSE
M00075	Catecholamine biosynthesis, tyrosine => dopamine => noradrenaline => adrenaline	2.1.1.28	10428	1	6	0.161593249	FALSE	FALSE
M00059*	Isoleucine degradation, isoleucine => propionyl-CoA	1.3.99.2	4203	1	8	0.257052496	FALSE	FALSE

Table S38: Gene losses in selected KEGG pathways. Pathways significantly enriched among lost genes are in **bold**. Pathways for which gene losses were manually validated using TBLASTN are labelled with an asterisk.

	Descriptions	**Lost genes from *Nasonia* (Entrez Gene ID of human ortholog)**	**Number of lost enzymes**	**Number of enzymes in the module**	**Enrichment p-value (by hypergeometric test)**	**Dunn-Sidak correction**
Pathway	MAPK signaling pathway	10125 10912 11221 1616 2002 2251 2261 4915 5320 5922 7186 785 80824	13	157	1.65E-002	FALSE
Pathway	Tryptophan metabolism*	125061 223 23498 267 29116 3033 7337 8942	8	49	5.14E-003	FALSE
Pathway	Oxidative phosphorylation	1537 4698 4702 4707 4710 528	6	85	6.33E-002	FALSE
Pathway	Cell cycle	10459 10912 23594 5001 51434 9088	6	85	5.91E-002	FALSE
Pathway	Wnt signaling pathway*	11211 1457 54361 6425 7472 85407 9350	7	104	0.288	FALSE
Pathway	Lysine degradation*	223 3033 55217 60559 6472 8424	6	39	4.13E-002	FALSE
Pathway	Butanoate metabolism	10994 223 3033 35 622 80821	6	31	7.33E-004	TRUE
Pathway	**Valine, leucine and isoleucine degradation**	223 3033 35 4594 84693	5	38	7.27E-003	FALSE
Pathway	**ABC transporters – General**	19 26154 5826 89845 9429	5	26	2.12E-003	TRUE
Pathway	**Glutathione metabolism**	2879 2944 2954 3418 4259	5	21	3.36E-003	TRUE

Table S39: Orthologous groups (OG) present in human and *Nasonia* but not other insects. The listed proteins have at least 1000x better TBLASTN hit with the *Nasonia* genome than to any of *T. castaneum*, *A. mellifera*, and *D. melanogaster.* Species abbreviations: *H. sap. = Homo sapiens*, *D. pul. = Daphnia pulex*, *P. hum. = Pediculus humanus*, *A. mel. = Amel mellifera*, *N. vit. = Nasonia vitripennis*, *T. cas. = Tribolium castaneum*, *D. mel. = Drosophila melanogaster*.

OG id	*H. sap.*	*D. pul.*	*P. hum.*	*A. mel.*	*N. vit.*	*T. cas.*	*D. mel.*	Human ortholog	Hugo name	Ensembl description
9567	1	0	0	0	1	0	0	ENSP00000306760	LRRC45	Leucine-rich repeat-containing protein 45. [Source:Uniprot/SWISSPROT;Acc:Q96CN5]
9182	1	2	0	0	1	0	0	ENSP00000340836	OSTF1	Osteoclast-stimulating factor 1. [Source:Uniprot/SWISSPROT;Acc:Q92882]
11607	1	0	0	0	2	0	0	ENSP00000370307	ASGR1	Asialoglycoprotein receptor 1 (ASGPR 1) (ASGP-R 1) (Hepatic lectin H1). [Source:Uniprot/SWISSPROT;Acc:P07306]
5688	1	1	0	0	1	0	0	ENSP00000362224	STK40	Serine/threonine-protein kinase 40 (EC 2.7.11.1) (SINK-homologous serine/threonine-protein kinase). [Source:Uniprot/SWISSPROT;Acc:Q8N2I9]
8816	1	1	0	0	1	0	0	ENSP00000298317	RPUSD4	RNA pseudouridylate synthase domain containing 4 [Source:RefSeq_peptide;Acc:NP_116184]
10951	1	1	0	0	1	0	0	ENSP00000353971	C8orf41	Uncharacterized protein C8orf41. [Source:Uniprot/SWISSPROT;Acc:Q6NXR4]
11473	1	0	0	0	1	0	0	ENSP00000349923	LAGE3	L antigen family member 3 (Protein ITBA2) (Protein ESO-3). [Source:Uniprot/SWISSPROT;Acc:Q14657]
5923	1	1	0	0	1	0	0	ENSP00000255078	IGHMBP2	DNA-binding protein SMUBP-2 (EC 3.6.1.-) (ATP-dependent helicase IGHMBP2) (Immunoglobulin mu-binding protein 2) (SMUBP-2) (Glial factor 1) (GF-1). [Source:Uniprot/SWISSPROT;Acc:P38935]
7661*	2	0	0	0	1	0	0	ENSP00000282146	KCNK13	Potassium channel subfamily K member 13 (Tandem pore domain halothane-inhibited potassium channel 1) (THIK-1). [Source:Uniprot/SWISSPROT;Acc:Q9HB14]
9242*	2	0	0	0	1	0	0	ENSP00000250024	E2F8	E2F family member 8 [Source:RefSeq_peptide;Acc:NP_078956]

* Both Human proteins are >1000x more similar with Nvit genome than to any other inse

Table S40: Runaway duplication in *Nasonia* of otherwise single copy genes. Number of paralogs are shown per orthologous group (OG) with human and fly orthologs. Putative function was inferred via the human ortholog. Species abbreviations: *H. sap. = Homo sapiens*, *D. pul. = Daphnia pulex*, *P. hum. = Pediculus humanus*, *A. mel. = Amel mellifera*, *N. vit. = Nasonia vitripennis*, *T. cas*. = *Tribolium castaneum*, *D. mel. = Drosophila melanogaster*.

OG id	*H. sap*	*D. pul*	*P. hum.*	*A. mel.*	*N. vit.*	*T. cas.*	*D. mel.*	Human ortholog	Flybase	Flybase name	Putative function
4545	1	1	1	1	6	1	1	ENSP00000305899	FBgn0025639	Suv4-20	Histone-lysine N-methyltransferase
1569	1	1	1	1	4	1	1	ENSP00000372499	FBgn0023444	ebi	F-box-like/WD repeat protein TBL1Y (Transducin beta-like 1Y protein). Regulation of epidermal growth factor receptor signaling pathway; Notch signaling pathway; epidermal growth factor receptor signaling pathway; regulation of proteolysis
4066	1	1	1	1	4	1	1	ENSP00000270637	FBgn0033375	CG8078	ATP-binding domain protein 3 (Cancer-associated gene protein)
2193	1	1	1	1	3	1	1	ENSP00000370031	FBgn0031255	BBS8	TTC8 Tetratricopeptide repeat protein 8. Function in ciliogenesis but is dispensable for centriolar satellite function
2332	1	1	1	1	3	1	1	ENSP00000261531	FBgn0004856	Bx42	SNW1 Nuclear receptor coactivator Involved in vitamin D-mediated transcription. Can function as a splicing factor in pre-mRNA splicing
2441	1	1	1	1	3	1	1	ENSP00000297161	FBgn0000395	crossveinless 2	Imaginal disc-derived wing vein specification; imaginal disc-derived wing morphogenesis.
3754	1	1	1	1	3	1	1	ENSP00000324343	FBgn0037513	pyd3	Beta-ureidopropionase (EC 3.5.1.6)(Beta-alanine synthase)(N-carbamoyl-beta-alanine amidohydrolase)(BUP-1) [Source:UniProtKB/Swiss-Prot;Acc:Q9UBR1]
528	1	1	1	1	3	1	1	ENSP00000355324	FBgn0086250	optic atrophy 1-like	OPA1 Dynamin-related GTPase required for mitochondrial fusion and regulation of apoptosis. May form a diffusion barrier for proteins stored in mitochondrial cristae. Proteolytic processing in response to intrinsic apoptotic signals may lead to disassembly of OPA1 oligomers and release of the caspase activator cytochrome C (CYCS) into the

											mitochondrial intermembrane space
6679	1	1	1	1	3	1	1	ENSP00000370480	FBgn0035206	CG9186	unknown
7703	1	1	1	1	3	1	1	ENSP00000261880	FBgn0052109	CG32109	Alpha- and gamma-adaptin-binding protein p34
8112	1	1	1	1	3	1	1	ENSP00000265881	FBgn0039115	CG10214	3'-to-5' exoribonuclease specific for small oligoribonucleotides. May have a role for cellular nucleotide recycling.

Table S41: Losses of otherwise single copy genes in *Nasonia* (TBLASTN hit of bee protein to *Nasonia* genome not better than 10^3). Single copy in five of *H. sapiens* (*H. sap.*), *D. pulex* (*D. pul.*), *P. humanus* (*P. hum.*), *A. mellifera* (*A. mel.*), *N.* vitripennis (*N.vit.*), *T. castaneum* (*T. cas.*), *D. melanogaster* (*D. mel.*). For each orthologous group (OG) the number of members is shown for each species.

OG id	*H. sap.*	*D. pul.*	*P. hum.*	*A. mel.*	*N. vit.*	*T. cas.*	*D. mel.*	E-value TBLASTN to *Nasonia* genome with bee protein	Bee	*T. cas.*	Fly	Flyname	Putative function (ensembl)
9937	0	1	1	1	0	1	1	1.30E-003	GB18635-PA	GLEAN_04532	FBgn0039336		
5543	1	1	1	1	0	1	1	1.60E-003	GB15289-PA	GLEAN_07615	FBgn0030655	CG4553	
8875	1	1	1	1	0	1	1	1.70E-003	GB17730-PA	GLEAN_07463	FBgn0032017	CG9213	CWF19
9439	2	1	1	1	0	1	1	2.60E-003	GB10163-PA	GLEAN_09107	FBgn0037490	CG7810	
9480	1	1	1	1	0	1	1	3.70E-003	GB30558-PA	GLEAN_00470	FBgn0033654	CG2249	
10563	1	1	0	1	0	1	1	5.50E-003	GB13171-PA	GLEAN_11887	FBgn0038860	Sobp	
5649	1	1	1	0	0	1	1	6.00E-003		GLEAN_04399	FBgn0033609	CG10825	
8523	1	0	1	1	0	1	1	6.30E-003	GB14355-PA	GLEAN_00839	FBgn0039155	CG13213	
11030	0	1	1	1	0	1	1	8.00E-003	GB12509-PA	GLEAN_06322	FBgn0051847	l(3)neo43	
9063	1	1	1	1	0	1	0	1.30E-002	GB10797-PA	GLEAN_13935		CG31847	
10222	1	1	1	0	0	1	1	1.50E-002		GLEAN_03252	FBgn0053052		
9207	2	1	1	1	0	1	1	1.50E-002	GB19282-PA	GLEAN_07480	FBgn0025626	CG33052	
4380	1	2	1	1	0	1	1	1.90E-002	GB12823-PA	GLEAN_14029	FBgn0051759	CG4957	
9205	1	1	1	1	0	1	0	2.30E-002	GB18356-PA	GLEAN_13984		CG17386	
4833	0	1	1	1	0	1	1	> 0.1	GB17678-PA	GLEAN_13541	FBgn0031257		
6198	1	1	1	1	0	1	1	> 0.1	GB16529-PA	GLEAN_09164	FBgn0034859	Tudor-SN	VIGILIN; ALTNAME: FULL=HIGH DENSITY LIPOPROTEIN BINDING PROTEIN; SHORT=HDL BINDING PROTEIN;
10661	1	1	1	1	0	1	0	> 0.1	GB13129-PA	GLEAN_04841		CG9007	#REF!
7408	1	1	0	1	0	1	1	> 0.1	GB12402-PA	GLEAN_15771	FBgn0033162	Mrtf	
8779	1	0	1	1	0	1	1	> 0.1	GB11001-PA	GLEAN_04927	FBgn0034803	Mtp	LACTOYLGLUTATHIONE LYASE; EC=4 4 1 5; ALTNAME: FULL=METHYLGLYOXALASE; ALTNAME: FULL=ALDOKETOMUTASE; ALTNAME: FULL=GLYOXALASE I; SHORT=GLX I; ALTNAME: FULL=KETONE ALDEHYDE MUTASE; ALTNAME: FULL=S D LACTOYLGLUTATHIONE METHYLGLYOXAL LYASE;
6025	1	1	1	1	0	1	1	> 0.1	GB12908-PA	GLEAN_05717	FBgn0031397	kz	PAP21 PROTEIN; FLAGS: PRECURSOR;
7483	2	1	1	1	0	1	1	> 0.1	GB19381-PA	GLEAN_12928	FBgn0033813	ss	
6012	1	1	1	1	0	1	1	> 0.1	GB16580-PA	GLEAN_11355	FBgn0037301	Cpr	

4228	1	1	1	1	0	1	1	> 0.1	GB10195-PA	GLEAN_06436	FBgn0001337	Vps16A	
4577	1	1	1	1	0	1	1	> 0.1	GB19583-PA	GLEAN_14215	FBgn0031812	CG9346	
2925	1	1	1	0	0	1	1	> 0.1		GLEAN_04359	FBgn0064766	APC7	FIC DOMAIN CONTAINING PROTEIN;
9730	1	1	1	1	0	1	1	> 0.1	GB15960-PA	GLEAN_08520	FBgn0051957	CG7207	
5156	1	1	1	1	0	1	1	> 0.1	GB19137-PA	GLEAN_08000	FBgn0027780	CG9425	PROBABLE RNA BINDING EIF1AD; ALTNAME: FULL=EUKARYOTIC TRANSLATION INITIATION FACTOR 1A DOMAIN CONTAINING PROTEIN;
2234	1	3	1	1	0	1	1	> 0.1	GB19185-PA	GLEAN_13924	FBgn0024689	Vha44	ACYL COA SYNTHETASE FAMILY MEMBER 4 HOMOLOG; EC=6 2 1 ;
9715	1	1	1	0	0	1	1	> 0.1		GLEAN_01712	FBgn0030559	CG5149	CONSERVED OLIGOMERIC GOLGI COMPLEX SUBUNIT 5; SHORT=COG COMPLEX SUBUNIT 5; ALTNAME: FULL=COMPONENT OF OLIGOMERIC GOLGI COMPLEX 5;
1916	1	1	1	1	0	1	1	> 0.1	GB15379-PA	GLEAN_02627	FBgn0029685	CG1530	
10877	1	1	1	1	0	1	1	> 0.1	GB13382-PA	GLEAN_11163	FBgn0052536		PROBABLE O ACETYLTRANSFERASE CAS1; EC=2 3 1 ; ALTNAME: FULL=CAPSULE SYNTHESIS 1;
5945	1	1	1	0	0	1	1	> 0.1		GLEAN_00193	FBgn0038045		
6754	1	0	1	1	0	1	1	> 0.1	GB15523-PA	GLEAN_12611	FBgn0034352	Rpn6	SIALIC ACID SYNTHASE; ALTNAME: FULL=N ACETYLNEURAMINATE SYNTHASE; EC=2 5 1 56; ALTNAME: FULL=N ACETYLNEURAMINIC ACID SYNTHASE; ALTNAME: FULL=N ACETYLNEURAMINATE 9 PHOSPHATE SYNTHASE; EC=2 5 1 57; ALTNAME: FULL=N ACETYLNEURAMINIC ACID PHOSPHA
7557	3	1	1	1	0	1	1	> 0.1	GB10223-PA	GLEAN_06172	FBgn0033807	CG8351	
3556	1	1	1	2	0	1	1	> 0.1	GB11367-PA	GLEAN_01186	FBgn0044028	CG31694	PROTEIN HRPAP20 HOMOLOG;
2602	1	1	1	1	0	0	1	> 0.1	GB10166-PA		FBgn0037347	Ahcy13	
10969	1	1	1	0	0	1	1	> 0.1		GLEAN_07114	FBgn0029937		O PHOSPHOSERYL TRNA SEC SELENIUM TRANSFERASE; EC=2 9 1 N1; ALTNAME: FULL=SELENOCYSTEINE SYNTHASE; SHORT=SEC SYNTHASE; ALTNAME: FULL=SELENOCYSTEINYL TRNA SEC SYNTHASE; ALTNAME: FULL=SEP TRNA:SEC TRNA SYNTHASE; SHORT=SEPSECS; ALTNAME: FULL=UGA
5718	1	1	1	0	0	1	1	> 0.1		GLEAN_11516	FBgn0029664	repo	
9974	1	1	1	1	0	1	1	> 0.1	GB17422-PA	GLEAN_12505	FBgn0038453	CG11281	ALANYL TRNA SYNTHETASE DOMAIN CONTAINING 1
7631	1	0	1	1	0	1	1	> 0.1	GB19326-PA	GLEAN_03894	FBgn0034925	CG8520	ARL 6 INTERACTING 1 HOMOLOG;
11526	0	1	1	1	0	1	1	> 0.1	GB15923-PA		FBgn0000500	CG8207	
8201	1	1	1	1	0	1	1	> 0.1	GB17290-PA	GLEAN_03140	FBgn0050493	mid	DROSULFAKININS; CONTAINS: RECNAME: FULL=DROSULFAKININ 0; SHORT=DSK 0;

												CONTAINS: RECNAME: FULL=DROSULFAKININ 1; ALTNAME: FULL=DROSULFAKININ I; SHORT=DSK I; CONTAINS: RECNAME: FULL=DROSULFAKININ 2; ALTNAME: FULL=DROSULFAKININ II; SHORT=DSK II;
10367	1	1	1	1	0	0	1	> 0.1	GB20083-PA		FBgn0029594	mus201	PA P ;
8764	1	1	1	0	0	1	1	> 0.1		GLEAN_15995	FBgn0052220	CkIIalpha	
7476	1	1	1	0	0	1	1	> 0.1		GLEAN_06374	FBgn0030799	Alp23B	
9630	1	1	1	0	0	1	1	> 0.1		GLEAN_13647	FBgn0038828	Rpb4	
8457	2	1	1	1	0	1	1	> 0.1	GB17996-PA	GLEAN_12817	FBgn0032703	Arc42	THIOREDOXIN DOMAIN CONTAINING 9; ALTNAME: FULL=ATP BINDING ASSOCIATED WITH CELL DIFFERENTIATION;
6879	0	1	1	1	0	1	1	> 0.1	GB16442-PA	GLEAN_04031	FBgn0035760	SF1	
5252	1	1	1	1	0	1	2	> 0.1	GB11145-PA	GLEAN_06652	FBgn0029949	CG6401	MITOCHONDRIAL; FLAGS: PRECURSOR;
5984	1	0	1	1	0	1	1	> 0.1	GB18319-PA	GLEAN_02934	FBgn0022960	Scsalpha	
8987	1	1	1	1	0	0	1	> 0.1	GB15418-PA		FBgn0037898	CG5706	
8179	1	1	1	1	0	1	1	> 0.1	GB15105-PA	GLEAN_04363	FBgn0051122	CG7414	D TYROSYL TRNA TYR DEACYLASE; EC=3 1 ;
2784	1	1	1	1	0	1	1	> 0.1	GB15630-PA	GLEAN_15929	FBgn0033225	CG3075	
8180	0	1	1	1	0	1	1	> 0.1	GB19786-PA	GLEAN_02618	FBgn0035755	CG4374	TUBULIN TYROSINE LIGASE 12;
10859	0	1	1	1	0	1	1	> 0.1	GB10338-PA	GLEAN_08329	FBgn0023526	CG4713	
5470	1	1	1	0	0	1	1	> 0.1		GLEAN_04932	FBgn0030520	CG13886	

Table S42: RT-PCR and 5'- and 3'-RACE-PCR primers used to study the transformer (*tra*) gene in *N. vitripennis*.

Application	Primer name	Primer sequence 5' – 3'	Exon
3'RACE	NvTra_F1	GAAAAGCGAAGGATTGCTTG	1
5'RACE	NvTra_R1	AGCAACTCGAAGCTTTTTGC	1
5'RACE	NvTra_R2	TTTCAGTTGCAGATGCAGGA	2
RT-PCR / 3'RACE	NvTra_F2	GACCAAAAGAGGCACCAAAA	2
RT-PCR / 5'RACE	NvTra_R3	GGCGCTCTTCCACTTCAAT	3

Table S44: Coordinates of the annotated genes encoding Yellow/Royal Jelly proteins in *Nasonia vitripennis.*

Gene name	Alternative gene name	NCBI GeneID	NCBI RefSeq	RefSeq status*	BCM Scaffold	NCBI Scaffold	NCBI RefSeq	Range	Strand
Mrjpl10	NvMRJPL10	100116124	XM_001600620.1	model	1011	NviUn_WGA1011_1	NW_001814584.1	44404..42674	minus
Yellow-f	Nv-yellow-f	100116091	NM_001161496.1	validated	1395	NviUn_WGA1395_1	NW_001815009.1	5880..841	minus
Mrjpl8	NvMRJPL8	100113745	NM_001161502.1	provisional	143	NviUn_WGA143_1	NW_001815059.1	163388..158141	minus
Mrjpl7	NvMRJPL7	100113779	NM_001161503.1	validated	143	NviUn_WGA143_1	NW_001815059.1	168827..166044	minus
Mrjpl6	NvMRJPL6	100113806	XM_001599068.1	model	143	NviUn_WGA143_1	NW_001815059.1	173154..171020	minus
Yellow-x1e	Nv-yellow-x1e	100123593	XM_001607202.1	model	16	NviUn_WGA16_1	NW_001815348.1	2761443..2763809	plus
Yellow-x1d	Nv-yellow-x1d	100123792	XM_001607460.1	model	16	NviUn_WGA16_1	NW_001815348.1	3239899..3238592	minus
Yellow-x2	Nv-yellow-x2	100116563	XM_001600972.1	model	21	NviUn_WGA21_1	NW_001815904.1	447174..452211	plus
Yellow	Nv-yellow	100114782	NM_001161505.1	provisional	36	NviUn_WGA36_1	NW_001817570.1	326773..317405	minus
Mrjpl9	NvMRJPL9	100117038	NM_001161506.1	validated	42	NviUn_WGA42_1	NW_001818237.1	Under revision	minus
Yellow-h	Nv-yellow-h	100119523	NM_001159922.1	provisional	42	NviUn_WGA42_1	NW_001818237.1	955592..950126	minus
Mrjpl5	NvMRJPL5	100119551	NM_001161507.1	validated	42	NviUn_WGA42_1	NW_001818237.1	959330..957627	minus
Mrjpl4	NvMRJPL4	100119591	NM_001161508.1	validated	42	NviUn_WGA42_1	NW_001818237.1	961691..960169	minus
Mrjpl3	NvMRJPL3	100119631	NM_001161509.1	validated	42	NviUn_WGA42_1	NW_001818237.1	xxxxxx..962179	minus
Mrjpl2	NvMRJPL2	100301998	NM_001161553.1	inferred	42	NviUn_WGA42_1	NW_001818237.1	966585..964994	minus
Mrjpl1	NvMRJPL1	100119671	XM_001603354.1	model	42	NviUn_WGA42_1	NW_001818237.1	968875..967191	minus
Yellow-e3	Nv-yellow-e3	100119705	NM_001161510.1	validated	42	NviUn_WGA42_1	NW_001818237.1	972658..969825	minus
Yellow-e	Nv-yellow-e	100119761	NM_001161513.1	provisional	42	NviUn_WGA42_1	NW_001818237.1	981160..978774	minus
Yellow-g2b	Nv-yellow-g2b	100119791	NM_001161514.1	validated	42	NviUn_WGA42_1	NW_001818237.1	992035..990772	minus
Yellow-g2a	Nv-yellow-g2a	100119831	XM_001603494.1	model	42	NviUn_WGA42_1	NW_001818237.1	995444..994071	minus
Yellow-g2c	Nv-yellow-g2c	100119867	XM_001603523.1	model	42	NviUn_WGA42_1	NW_001818237.1	997864..996577	minus
Yellow-g	Nv-yellow-g	100119898	XM_001603551.1	model	42	NviUn_WGA42_1	NW_001818237.1	998768..1000069	plus
Yellow-b	Nv-yellow-b	100120303	NM_001161517.1	validated	5	NviUn_WGA5_1	NW_001820126.1	3480122..3487302	plus
Yellow-x1c	Nv-yellow-x1c	100114787	NM_001161518.1	validated	62	NviUn_WGA62_1	NW_001820338.1	103800..105282	plus
Yellow-x1a	Nv-yellow-x1a	100117576	XM_001601721.1	model	62	NviUn_WGA62_1	NW_001820338.1	714304..715735	plus
Yellow-x1b	Nv-yellow-x1b	100117619	NM_001161554.1	inferred	62	NviUn_WGA62_1	NW_001820338.1	715867..719302	plus

* the status of the record definitions are explained at http://www.ncbi.nlm.nih.gov/RefSeq/key.html#status

Table S46: List of epigenetic control related genes with different copy numbers between 4 insect species. Among four insect species examined, *Drosophila* has increased the copy number of genes encoding five proteins associated with epigenetic control.

Protein name	Biochemical complex/function	*Drosophila*	*Tribolium*	*Apis*	*Nasonia*
Extra sex combs	PRC2 subunit	2	1	1	1
Pleiohomeotic	DNA binding, PcG recruiter	2	1	1	1
Zeste	DNA binding, TrxG/PcG recruiter	1	0	0	0
JMJD	H3K36 demethylase	2	1	1	1
HP1	Chromatin binding through H3K9	4	3	2	2

Table S47: Gene structure statistics for *Nasonia* in comparison to other eukaryotes, tabulated at http://insects.eugenes.org/arthropods/data/summaries/.

	Bee	Wasp[1]	Beetle	Mosquito	Fruitfly[1]	Mouse	Worm
Genome size (Mb)	220	290	180	580	180	3,450	100
	(200)	(250)	(177)	(400)	(120)	(2,600)	(100)
No. of genes	17,000	27,300	16,400	18,900	13,700	20,600	20,100
Gene density	0.040	0.120	0.100	0.055	0.168	0.015	0.250
Gene length	3,900	2,900	3,700	3,700	3,200	8,300	3,000
CDS size	1,590	1,510	1,370	1,280	1,650	1,820	1,300
Exons/gene	6.8	6.0	4.4	3.5	4.3	6.3	6.0
Exon size [2]	240	260	310	360	410	280	200
Intron size [3]	79/290	79/310	57/1200	65/1100	69/400	1200/90	65/400
Mean	770	430	1000	1600	660	2800	290
Intr > Exon	36%	24%	31%	37%	27%	85%	33%
UTR size [4]	340	680	--	190	750	--	260
Intergenic size	9,200	--	7,900	17,500	4,700	68,000	2,400

[1] Gene part sizes and exons/gene are measured with EST-validated gene models for these noted genomes. Others are measured from reference database gene data. [2] Exon size distribution for Fruitfly is strongly bimodal, with single-exon genes twice the size of multi-exon genes. Other species show a more unimodal distribution of exon sizes. [3] Intron size is non-normally distributed. This row lists primary and secondary peaks, mean and percent of introns larger than exons. [4] UTR size measures only exons that extend past coding sequence, missing true cases of zero length UTRs. Species listed are Beetle = *Tribolium castenatum* (tcas3); Bee = *Apis mellifera* (ncbi1); Fruitfly = *Drosophila melanogaster* (fb5.5); Mosquito = *Culex pipens* (cpip12); Mouse = *Mus musculus* (mgi3); Wasp = *Nasonia vitripennis* (nvit1); Worm = *Caen. elegans* (wb167).

Table S48: Arthropod proteome sources for gene clustering. Number of protein coding genes (N genes), transcripts (N trans.) and genome size in megabases are given.

Species	N genes	N trans.	Genome size	Source*
Aedes aegyptii	15419	16789	1310	VectorBase AaegL1.1, June 2006
Anopheles gambiae	13538	14281	278	VectorBase AgamP3, Feb. 2006
Culex pipiens	18883	18883	579	VectorBase CpipJ1.2, June 2008
Drosophila melanogaster	13421	20507	169	FlyBase release 5.5, Feb 2008
Drosophila pseudoobscura	18910	19259	153	NCBI Gnomon [G] CAF1, May 2006
Drosophila mojavensis	17785	17950	194	NCBI Gnomon CAF1, May 2006
Tribolium castaneum	16344	16422	176	Beetlebase release 3, Mar 2008
Apis mellifera	17025	17182	217	NCBI Gnomon rel. 4, Aug 2006
Nasonia vitripennis	27203	27287	294	NCBI Gnomon, Dec 2007
Pediculus humanus	11194	11198	112	VectorBase PhumU1.1, April 2007
Acyrthosiphon pisum	37783	37994	464	NCBI Gnomon, June 2008
Daphnia pulex	37329	37466	227	NCBI Gnomon, wFleaBase.org, May 2007
Ixodes scapularis	17742	17742	1765	VectorBase IscaW1.0.5, Apr 2008

* Arthropod gene sources accessed on July 2008 are NCBI Gnomon gene predictions, described in SOM Material and Methods (NCBI's gene prediction pipeline), VectorBase gene sets described at http://www.vectorbase.org/, *Tribolium* gene set from http://beetlebase.org/, and *Drosophila melanogaster* gene set from http://flybase.org/

Table S49: Arthropod gene model count, partitioned by orthology and paralogy.

Species	nGene	TE	Singletons		Duplicates		Relative to Dipterans	
			Uniq1	Orth1	UDup	OrDup	Single	Double
Nasonia	26703	7569	6592	7317	3191	2034	1.1	1.5
Apis	15605	53	5627	7595	624	1706	1.1	0.7
Tribolium	16284	285	5323	7367	1359	1950	1	1
DrosPse	18610	362	4337	9955	1804	2152		
DrosMel	13956	312	1429	10623	443	1325	1	1
Culex	18851	152	4008	9075	1954	3662		
Aedes	15415	168	1922	8700	775	3906		

Gene count categories: nGene = total gene predictions used, Uniq1 = unique singleton (no ortholog or paralog), UniqDup = species paralogs (no ortholog), Orth1 = has ortholog but no paralog , OrthDup = has ortholog and paralogs, TE = estimated transposon genes. nGene is lower than source genes as short (< 40 aa) models were removed before analysis. Gene analyses are available at arthropod gene group reports, http://insects.eugenes.org/arthropods/ (2008) (see also Gilbert 2009).

References:

Gilbert DG (2009): Aphid and Waterflea have a High Rate of Gene Duplications Compared to Other Arthropods. PLoS ONE, provisionally accepted, May 2009.

Table S51: Overabundant *Nasonia* gene families, compared to insects. Columns are Nas = *Nasonia* genes in group, Iave = average number of genes among nine other insect species, Imax = maximum among nine other insects. Gene Groups are reported in detail at the arthropod gene group reports, http://insects.eugenes.org/arthropods/ (2008) (see also Gilbert 2009), and insect species gene sets are given in Table S2.

Gene group	Nas	IAve	IMax	Description
ARP1_G7351	9	0.1	1	Acid phosphatase-1
ARP1_G5927	8	0.4	1	Alcohol dehydrogenase
ARP1_G9938	5	0.1	1	Amine oxidase
ARP1_G12568	3	0.1	1	Apis mellifera protein (LOC726903)
ARP1_G6777	6	0.5	5	Apis mellifera protein (LOC727283)
ARP1_G7382	9	0	0	Apocytochrome b
ARP1_G10249	5	0.1	1	ATP synthase subunit 6
ARP1_G3370	12	0.2	1	Chemosensory receptor 9
ARP1_G822	8	1.1	6	Coumarate-CoA ligase 1
ARP1_G12485	3	0.1	1	Cytochrome P450
ARP1_G250	18	2	8	Cytochrome P450 4C1
ARP1_G9784	4	0.2	1	Deoxyribonuclease I
ARP1_G406	11	1.6	5	Glucose dehydrogenase
ARP1_G242	8	1.6	5	Glucosyl/glucuronosyl transferases
ARP1_G9915	5	0.1	1	Glucosyl/glucuronosyl transferases
ARP1_G3335	7	0.7	2	Juvenile hormone-inducible protein
ARP1_G567	9	1	1	Laccase 1 (EC 1.10.3.2)
ARP1_G12565	3	0.1	1	Larval serum protein 2
ARP1_G1836	15	0	0	Lectin, 26-kDa
ARP1_G9928	5	0.1	1	Leucine-rich transmembrane protein
ARP1_G11446	4	0	0	Lysosomal acid lipase
ARP1_G12517	3	0.1	1	Male ejaculatory bulb protein III
ARP1_G9948	5	0.1	1	Male sterility protein 2
ARP1_G8509	7	0.1	1	Metalloprotease
ARP1_G336	10	0.3	1	Microtubule binding protein
ARP1_G7902	7	0.2	1	MPA3 allergen
ARP1_G9117	6	0.1	1	NADH dehydrogenase subunit 5
ARP1_G285	8	1.9	6	Neprilysin 2
ARP1_G288	30	0.4	4	Odorant receptor
ARP1_G548	23	0.1	1	Odorant receptor 30a
ARP1_G5928	11	0.1	1	odorant receptor 47
ARP1_G72	15	5.6	15	Oxoacyl- reductase 1
ARP1_G10923	4	0.2	2	Pancreatic lipase-related protein 2 precursor
ARP1_G7771	5	0.4	2	Pupal cuticle protein
ARP1_G11447	4	0	0	Ribonuclease NW
ARP1_G7409	6	0.4	1	Serine protease 2
ARP1_G486	8	1.7	6	Sodium-dependent phosphate transporter
ARP1_G1402	8	0.9	5	Sodium/solute symporter
ARP1_G8602	6	0.2	2	Tyrosine recombinase
ARP1_G12520	3	0.1	1	Venom acid phosphatase

References:

Gilbert DG (2009): Aphid and Waterflea have a High Rate of Gene Duplications Compared to Other Arthropods. PLoS ONE, provisionally accepted, May 2009.

Table S52: Evolutionary rates for *Nasonia* gene classes.

Gene category	Statistic	Mean	Median	N	p-value
All RefSeq	dN/dS	0.2567	0.1736	7292	
	dN	0.0069	0.0051		
	dS	0.0318	0.0295		
Heterochromatic	dN/dS	0.3291	0.2121	853	$<2x10^{-5}$
	dN	0.0081	0.0063		
	dS	0.0326	0.0300		
Euchromatic	dN/dS	0.2471	0.1693	6432	$<2x10^{-5}$
	dN	0.0067	0.0050		
	dS	0.0317	0.0295		
Venom	dN/dS	0.5561	0.3624	43	$<2x10^{-6}$
	dN	0.0164	0.0119		
	dS	0.0344	0.0328		
Nasonia-specific	dN/dS	0.4723	0.2879	434	$<7x10^{-6}$
	dN	0.0122	0.0101		$<8x10^{-6}$
	dS	0.0385	0.0341		$<8x10^{-6}$
Hymenoptera-specific	dN/dS	0.3318	0.2532	147	$<2x10^{-6}$
	dN	0.0093	0.0080		$<2x10^{-6}$
	dS	0.0342	0.0328		<0.003

Table S53: Distribution of dN/dS across gene sizes for *Nasonia*-specific, Hymenoptera-specific, and non-specific genes.

Gene Size	Category	Number with dN/dS > 1	Number with dN/dS < 1	Percent with dN/dS > 1	Percent in size category
<500 bp	Hymenoptera-specific	2	14	13	11
	Nasonia-specific	25	203	11	53
	Non-specific	24	572	4	9
500-1000 bp	Hymenoptera-specific	2	27	7	20
	Nasonia-specific	13	104	11	27
	Non-specific	70	1694	4	26
>1000 bp	Hymenoptera-specific	3	99	3	69
	Nasonia-specific	5	84	6	21
	Non-specific	55	4359	1	65
All	Hymenoptera-specific	7	140	5	
	Nasonia-specific	43	391	10	
	Non-specific	149	6566	2	

Table S54: Genomic organization of detoxification genes. **(A)** CCE. **(B)** GST. **(C)** P450.
(A)

Gene name	Recombination	Gene density	RT content	Clade	OGS v1.2	CDS_start	Marker	cM	Scaffold	RefSeq
Nv1607750	--	high	low	I	XM_001607700.1	3721388	1,11	22,1	SCAFFOLD16	NW_001815348.1
Nv1604042	--	low	high	A	XM_001603992.1	1705535	1,21	45,3	SCAFFOLD12	NW_001814904.1
Nv1599255	low	low	high	F	XM_001599205.1	14070	1,24	48	SCAFFOLD74	NW_001820471.1
Nv1602248	low	low	high	B/C	XM_001602198.1	749233	1,24	48	SCAFFOLD61	NW_001820327.1
Nv1602279	low	low	high	B/C	XM_001602229.1	752856	1,24	48	SCAFFOLD61	NW_001820327.1
Nv1602306	low	low	high	B/C	XM_001602256.1	758013	1,24	48	SCAFFOLD61	NW_001820327.1
Nv1603146	low	low	high	F	XM_001603096.1	120467	1,25	48,9	SCAFFOLD60	NW_001820237.1
Nv1602331	low	low	high	A	XM_001602281.1	284169	1,26	49,8	SCAFFOLD58	NW_001820014.1
Nv1602356	low	low	high	A	XM_001602306.1	288892	1,26	49,8	SCAFFOLD58	NW_001820014.1
Nv1602413	low	low	high	A	XM_001602363.1	298810	1,26	49,8	SCAFFOLD58	NW_001820014.1
Nv1603773	low	low	high	E	XM_001603723.1	808470	1,26	49,8	SCAFFOLD58	NW_001820014.1
Nv1599913	low	low	high	E	XM_001599863.1	227963	2,28	43,3	SCAFFOLD76	NW_001820493.1
Nv1602765	low	low	high	E	XM_001602715.1	206803	2,28	43,3	SCAFFOLD79	NW_001820526.1
Nv1603087	low	low	high	E	XM_001603037.1	368038	2,28	43,3	SCAFFOLD79	NW_001820526.1
Nv1603114	low	low	high	E	XM_001603064.1	371828	2,28	43,3	SCAFFOLD79	NW_001820526.1
Nv1601401	low	low	high	B/C	XM_001601351.1	207494	2,29	44,2	SCAFFOLD179	NW_001815458.1
Nv1602389	low	low	high	E	XM_001602339.1	1575340	2,29	44,2	SCAFFOLD11	NW_001814793.1
Nv1604356	low	low	high	E	XM_001604306.1	1571102	2,29	44,2	SCAFFOLD11	NW_001814793.1
Nv1605919	low	low	high	E	XM_001605869.1	2337766	2,3	45,1	SCAFFOLD13	NW_001815015.1
Nv1605936	low	low	high	E	XM_001605886.1	2347217	2,3	45,1	SCAFFOLD13	NW_001815015.1
Nv1605713	--	low	low	H	XM_001605663.1	2088970	2,31	46	SCAFFOLD13	NW_001815015.1
Nv1605341	--	low	low	A	XM_001605291.1	1543124	3,05	9,3	SCAFFOLD18	NW_001815570.1
Nv1599778	low	high	high	E	XM_001599728.1	111437	3,26	46,1	SCAFFOLD89	NW_001820637.1
Nv1599809	low	high	high	E	XM_001599759.1	118222	3,26	46,1	SCAFFOLD89	NW_001820637.1
Nv1603542	low	low	high	B/C	XM_001603492.1	623977	3,27	47	SCAFFOLD37	NW_001817681.1
Nv1605568	--	low	low	J	XM_001605518.1	1033202	3,31	51,5	SCAFFOLD8	NW_001820638.1
Nv1604741	low	low	high	L	XM_001604691.1	805797	4,22	50,1	SCAFFOLD43	NW_001818348.1
Nv1604789	low	low	high	L	XM_001604739.1	827941	4,22	50,1	SCAFFOLD43	NW_001818348.1
Nv1606858	--	high	low	L	XM_001606808.1	3242769	4,33	73,6	SCAFFOLD9	NW_001820749.1
Nv1600458	--	high	low	J	XM_001600408.1	203076	4,41	87,6	SCAFFOLD9	NW_001820749.1
Nv1608086	--	low	high	A	XM_001608036.1	8117273	5,23	34	SCAFFOLD1	NW_001815682.1
Nv1608088	--	low	high	A	XM_001608038.1	8121352	5,23	34	SCAFFOLD1	NW_001815682.1
Nv1608200	--	low	high	A	XM_001608150.1	9154311	5,25	35,8	SCAFFOLD1	NW_001815682.1
Nv1603584	--	low	high	D	XM_001603534.1	1305552	5,32	42,2	SCAFFOLD10	NW_001814682.1

Nv1604190	--	low	low	K	XM_001604140.1	1715901	5,46	66,9	SCAFFOLD2	NW_001816793.1
Nv1601375	--	high	low	D	XM_001601325.1	765446	5,5	72,4	SCAFFOLD2	NW_001816793.1
Nv1601317	--	high	low	D	XM_001601267.1	748562	5,51	76,1	SCAFFOLD2	NW_001816793.1
Nv1601350	--	low	low	D	XM_001601300.1	752541	5,51	76,1	SCAFFOLD2	NW_001816793.1

(B)

Gene name	Recombination	Gene density	RT content	Clade	OGSv1.2	CDS_start	Marker	cM	Scaffold	RefSeq
NvGSTS1	--	high	low	Sigma	XM_001608175.1	3767760	1,11	22,1	SCAFFOLD16	NW_001815348.1
NvGSTS2	--	high	low	Sigma	XM_001607774.1	3891575	1,12	26,8	SCAFFOLD16	NW_001815348.1
NvGSTD4	low	low	high	Delta	XM_001600137.1	845437	1,24	48	SCAFFOLD32	NW_001817126.1
NvGSTO2	--	high	low	Omega	XM_001600982.1	361171	1,59	92,7	SCAFFOLD7	NW_001820527.1
NvGSTO1	--	high	low	Omega	XM_001600712.1	255885	1,6	93,6	SCAFFOLD7	NW_001820527.1
NvGSTT1	--	high	low	Theta	XM_001603636.1	261570	2,02	0,9	SCAFFOLD8	NW_001820638.1
NvGSTT2	--	high	low	Theta	XM_001605253.1	263353	2,02	0,9	SCAFFOLD8	NW_001820638.1
NvGSTT3	--	high	low	Theta	XM_001603664.1	264493	2,02	0,9	SCAFFOLD8	NW_001820638.1
NvGSTS7	low	low	high	Sigma	XM_001600423.1	67625	3,27	47	SCAFFOLD30	NW_001816904.1
NvGSTS8	low	low	high	Sigma	XM_001600927.1	104830	3,27	47	SCAFFOLD30	NW_001816904.1
NvGSTS3	--	high	low	Sigma	XM_001599361.1	24869	3,31	51,5	SCAFFOLD22	NW_001816015.1
NvGSTS4	--	low	low	Sigma	XM_001603892.1	2000234	3,35	56,1	SCAFFOLD22	NW_001816015.1
NvGSTS5	--	high	low	Sigma	XM_001605406.1	1060401	3,5	91,2	SCAFFOLD28	NW_001816681.1
NvGSTS6	--	high	low	Sigma	XM_001605048.1	1058616	3,5	91,2	SCAFFOLD28	NW_001816681.1
NvGSTD1	--	high	low	Delta	XM_001607805.1	2563411	5,13	17,4	SCAFFOLD7	NW_001820527.1
NvGSTD3	--	high	low	Delta	XM_001606124.1	2559062	5,13	17,4	SCAFFOLD7	NW_001820527.1
NvGSTD5	low	low	high	Delta	XM_001608157.1	9369681	5,26	36,8	SCAFFOLD1	NW_001815682.1

(C)

Gene name	Recombination	Gene density	RT content	Clade	OGSv1.2	CDS_start	Marker	cM	Scaffold	RefSeq
CYP9AG5	--	high	low	9	XM_001603807.1	588025	1,02	1,8	SCAFFOLD16	NW_001815348.1
CYP4AB16	--	high	low	4	XM_001606689.1	2279083	1,05	10,2	SCAFFOLD16	NW_001815348.1
CYP4AB17P	--	high	low	4	XM_001606604.1	2234814	1,05	10,2	SCAFFOLD16	NW_001815348.1
CYP9AH6	--	low	low	9	XM_001605022.1	674200	1,16	37,9	SCAFFOLD127	NW_001814881.1
CYP4AB6	--	high	low	4	XM_001604484.1	399731	1,17	41,7	SCAFFOLD127	NW_001814881.1
CYP9AH1	--	low	low	9	XM_001605011.1	2116759	1,18	42,6	SCAFFOLD24	NW_001816237.1
CYP9AH2	--	low	low	9	XM_001605030.1	2118194	1,18	42,6	SCAFFOLD24	NW_001816237.1
CYP9AH5	--	low	low	9	XM_001605066.1	2125867	1,18	42,6	SCAFFOLD24	NW_001816237.1
CYP336C1	low	low	low	??	XM_001599727.1	57254	1,22	46,2	SCAFFOLD87	NW_001820615.1
CYP6AS29	low	low	low	6	XM_001600233.1	366072	1,22	46,2	SCAFFOLD87	NW_001820615.1
CYP305A1	low	low	high	??	XM_001599989.1	503830	1,23	47,1	SCAFFOLD36	NW_001817570.1
CYP4G44	low	low	high	4	XM_001600251.1	467663	1,23	47,1	SCAFFOLD36	NW_001817570.1
CYP6AS37P	low	low	high	6	XM_001599779.1	348034	1,23	47,1	SCAFFOLD36	NW_001817570.1
CYP6BC2	low	low	high	6	XM_001599164.1	195468	1,23	47,1	SCAFFOLD36	NW_001817570.1
CYP15A1	low	low	high	??	XM_001605535.1	1034993	1,24	48	SCAFFOLD47	NW_001818792.1

CYP303A1	low	low	high	??	XM_001605158.1	403061	1,24	48	SCAFFOLD39	NW_001817903.1
CYP305D1	low	low	high	??	XM_001605550.1	1042205	1,24	48	SCAFFOLD47	NW_001818792.1
CYP334A1	low	low	high	??	XM_001599481.1	322046	1,24	48	SCAFFOLD32	NW_001817126.1
CYP6AQ7	low	low	high	6	XM_001599311.1	81916	1,24	48	SCAFFOLD164	NW_001815292.1
CYP6AQ9	low	low	high	6	XM_001602602.1	105536	1,24	48	SCAFFOLD149	NW_001815125.1
CYP6AS34	low	low	high	6	XM_001603688.1	164080	1,24	48	SCAFFOLD348	NW_001817336.1
CYP6AS35	low	low	high	6	XM_001603688.1	164080	1,24	48	SCAFFOLD348	NW_001817336.1
CYP6BD3	low	low	high	6	XM_001599410.1	107007	1,24	48	SCAFFOLD164	NW_001815292.1
CYP6CL1	low	low	high	6	XM_001603713.1	172879	1,24	48	SCAFFOLD348	NW_001817336.1
CYP6CK7	low	low	high	6	XM_001603096.1	120467	1,25	48,9	SCAFFOLD60	NW_001820237.1
CYP6CK6	low	low	high	6	XM_001602331.1	295312	1,26	49,8	SCAFFOLD58	NW_001820014.1
CYP6AS31	--	low	low	6	XM_001601609.1	188126	1,27	51,6	SCAFFOLD83	NW_001820571.1
CYP9AG8P	--	low	low	9	XM_001600010.1	312468	1,29	53,4	SCAFFOLD33	NW_001817237.1
CYP4AB23	--	high	low	4	XM_001607678.1	5707796	1,58	89,9	SCAFFOLD1	NW_001815682.1
CYP6BD5	--	low	low	6	XM_001603458.1	1748199	2,23	35,9	SCAFFOLD31	NW_001817015.1
CYP315A1	--	low	low	??	XM_001607584.1	4357164	2,25	37,7	SCAFFOLD5	NW_001820126.1
CYP336B1	low	low	high	??	XM_001603916.1	588505	2,28	43,3	SCAFFOLD15	NW_001815237.1
CYP4AB15	low	low	high	4	XM_001605343.1	1658400	2,28	43,3	SCAFFOLD15	NW_001815237.1
CYP4BW3	low	low	high	4	XM_001604018.1	627838	2,28	43,3	SCAFFOLD15	NW_001815237.1
CYP4BW4	low	low	high	4	XM_001603994.1	623238	2,28	43,3	SCAFFOLD15	NW_001815237.1
CYP4BW5	low	low	high	4	XM_001603967.1	617327	2,28	43,3	SCAFFOLD15	NW_001815237.1
CYP4AB14	low	low	high	4	XM_001606259.1	2836713	2,29	44,2	SCAFFOLD13	NW_001815015.1
CYP4BZ2	low	low	high	4	XM_001603165.1	1924760	2,29	44,2	SCAFFOLD11	NW_001814793.1
CYP6AQ4	low	low	high	6	XM_001606676.1	2819931	2,29	44,2	SCAFFOLD11	NW_001814793.1
CYP6AQ5	low	low	high	6	XM_001604675.1	2824453	2,29	44,2	SCAFFOLD11	NW_001814793.1
CYP4CA1	--	low	low	4	XM_001602929.1	303708	2,32	47,9	SCAFFOLD13	NW_001815015.1
CYP4BZ1	--	low	low	4	XM_001602345.1	361959	2,34	51,5	SCAFFOLD19	NW_001815681.1
CYP6AQ8	--	high	low	6	XM_001602297.1	575970	2,55	85,2	SCAFFOLD24	NW_001816237.1
CYP6CK8	--	high	high	6	XM_001606199.1	2559502	3,09	14,8	SCAFFOLD18	NW_001815570.1
CYP6CK9P	--	high	high	6	XM_001606190.1	2557198	3,09	14,8	SCAFFOLD18	NW_001815570.1
CYP6CK4	--	low	low	6	XM_001606993.1	3804256	3,17	31,4	SCAFFOLD6	NW_001820416.1
CYP4AB13	--	low	low	4	XM_001603661.1	919775	3,25	44,3	SCAFFOLD6	NW_001820416.1
CYP6CK10	low	low	high	6	XM_001601755.1	778260	3,27	47	SCAFFOLD89	NW_001820637.1
CYP6CK11	low	low	high	6	XM_001601782.1	783711	3,27	47	SCAFFOLD89	NW_001820637.1
CYP6CK12	low	low	high	6	XM_001602737.1	41745	3,27	47	SCAFFOLD44	NW_001818459.1
CYP4C60	low	low	high	4	XM_001603426.1	1854409	3,28	47,9	SCAFFOLD20	NW_001815793.1
CYP4AB18	--	low	low	4	XM_001606836.1	3523484	3,3	49,7	SCAFFOLD8	NW_001820638.1
CYP4AB12	--	low	low	4	XM_001601836.1	1181281	3,33	53,3	SCAFFOLD22	NW_001816015.1

CYP4G43	--	low	low	4	XM_001606367.1	2104697	3,45	79,2	SCAFFOLD17	NW_001815459.1
CYP4AA1	--	low	low	4	XM_001602061.1	25704	3,47	84,8	SCAFFOLD28	NW_001816681.1
CYP4AB7	--	high	low	4	XM_001604498.1	891019	3,49	89,4	SCAFFOLD28	NW_001816681.1
CYP4AB8	--	high	low	4	XM_001604475.1	887603	3,49	89,4	SCAFFOLD28	NW_001816681.1
CYP301A1	--	high	low	??	XM_001605622.1	1252866	3,5	91,2	SCAFFOLD28	NW_001816681.1
CYP301B1	--	high	low	??	XM_001605637.1	1263104	3,5	91,2	SCAFFOLD28	NW_001816681.1
CYP4AV5	--	low	low	4	XM_001607135.1	5121737	4,01	0	SCAFFOLD4	NW_001819015.1
CYP4AB4	--	high	low	4	XM_001606215.1	3739500	4,04	5,6	SCAFFOLD4	NW_001819015.1
CYP4AB5	--	high	low	4	XM_001606207.1	3736179	4,04	5,6	SCAFFOLD4	NW_001819015.1
CYP307A1	low	low	high	??	XM_001603385.1	277979	4,2	47,2	SCAFFOLD52	NW_001819348.1
CYP4AB10	low	low	high	4	XM_001601772.1	787863	4,2	47,2	SCAFFOLD29	NW_001816792.1
CYP4AB11	low	low	high	4	XM_001601772.1	787863	4,2	47,2	SCAFFOLD29	NW_001816792.1
CYP4AB9	low	low	high	4	XM_001601802.1	792249	4,2	47,2	SCAFFOLD29	NW_001816792.1
CYP314A1	low	low	high	??	XM_001599974.1	129916	4,21	48,2	SCAFFOLD123	NW_001814837.1
CYP9AG6	--	high	low	9	XM_001603961.1	1809569	4,38	82	SCAFFOLD9	NW_001820749.1
CYP12K1	--	low	low	??	XM_001604760.1	926745	5,11	15,6	SCAFFOLD38	NW_001817792.1
CYP336-un1	--	low	low	??	XM_001607071.1	3392458	5,12	16,5	SCAFFOLD7	NW_001820527.1
CYP6BD2	--	low	low	6	XM_001606940.1	3132751	5,12	16,5	SCAFFOLD7	NW_001820527.1
CYP6CK5	--	low	low	6	XM_001606637.1	2903402	5,12	16,5	SCAFFOLD7	NW_001820527.1
CYP6BD4				6	XM_001605790.1	2362902	5,14	19,3	SCAFFOLD7	NW_001820527.1
CYP6AS30	--	low	low	6	XM_001604171.1	1571669	5,17	26,7	SCAFFOLD7	NW_001820527.1
CYP4AB20	--	low	low	4	XM_001599684.1	336098	5,29	39,5	SCAFFOLD110	NW_001814693.1
CYP6AS27	--	low	low	6	XM_001604772.1	1803990	5,34	44	SCAFFOLD10	NW_001814682.1
CYP9AG7	--	low	low	9	XM_001601697.1	815	5,35	45,8	SCAFFOLD112	NW_001814715.1
CYP9P4	--	low	low	9	XM_001605979.1	2416030	5,35	45,8	SCAFFOLD10	NW_001814682.1
CYP9P5	--	low	low	9	XM_001605990.1	2418648	5,35	45,8	SCAFFOLD10	NW_001814682.1
CYP9AG1	--	low	high	9	XM_001607966.1	7526748	5,36	46,7	SCAFFOLD2	NW_001816793.1
CYP9AG2	--	low	high	9	XM_001607970.1	7530152	5,36	46,7	SCAFFOLD2	NW_001816793.1
CYP9AH7	--	low	high	9	XM_001607972.1	7539917	5,36	46,7	SCAFFOLD2	NW_001816793.1
CYP9AG3	--	high	low	9	XM_001601393.1	774603	5,5	72,4	SCAFFOLD2	NW_001816793.1
CYP9AG4	--	high	low	9	XM_001600636.1	771855	5,5	72,4	SCAFFOLD2	NW_001816793.1

Table S56: Female and male expression variation in response to *Wolbachia* infection. Expression variation data for each gene are based on quantification of gene transcripts from *Wolbachia*-infected and *Wolbachia*-uninfected stocks from tetracycline-cured lines. Levels of *Nasonia* immunity gene transcripts were normalized with the expression of the *Nasonia* ribosomal protein 49 gene transcripts. Data are means ± standard deviations of two independent replicates conducted with duplicate reactions for both males and females.

Immunity Gene	Female (mean ± SD)	Male (mean ± SD)
c-*jun* protein	2.69 ± 1.52	0.89 ± 0.57
spatzle-1B	1.81 ± 0.03	1.32 ± 0.60
myd88	1.04 ± 0.51	2.23 ± 2.07
relish	1.28 ± 0.50	1.54 ± 1.34
capsase-1	1.06 ± 0.44	-
tube	1.54 ± 0.88	0.92 ± 0.13
cactus	0.76 ± 0.47	1.40 ± 0.82
TEP III	2.43 ± 2.50	1.07 ± 0.65
PGRP-LC domain 10	1.43 ± 1.10	1.52 ± 1.16
Glucan recognition protein	1.47 ± 0.56	0.78 ± 0.36
PGRP-LC domain 7	0.58 ± 0.27	0.62 ± 0.01

Table S57: QTL for male courtship behavior. LOD = log odds ratio. Two QTL for the "number of headnods" in cycle 1and 2 were detected on chromosome 1 and 5 but only one QTL on chromosome 1 for cycle 3 and 4. All QTL for "number of headnods" explained more than 10 % of the additive variance observed in our mapping population. For "total number of cycles" we found 2 QTL, one on chromosome 2 and one on chromosome 4. Both together explained about 20 % of the additive variance observed in our mapping population. Genome wide significance thresholds were determined using a standard permutation test.

QTL location	LOD	% Explained Additive Variance	Marker
Headnod 1st cycle, (LOD 2.8 is the genome wide threshold[1])			
Chromosome 1	4.58	14.6	Scaf127_676512
Chromosome 5	3.14	10.5	Scaf176_257755
Headnod 2nd cycle, (LOD 2.8 is the genome wide threshold[1])			
Chromosome 1	2.98	12.4	Scaf127_676512
Chromosome 5	2.91	10.1	Scaf176_257755
Headnod 3rd cycle, (LOD 2.8 is the genome wide threshold[1])			
Chromosome 1	2.91	10.6	Scaf1_3410194
Headnod 4th cycle, (LOD 2.8 is the genome wide threshold[1])			
Chromosome 1	3.13	15.7	Scaf127_676512
Number of cycles, (LOD 2.3 is the genome wide threshold[1])			
Chromosome 2	2.54	8.5	Scaf116_395861
Chromosome 4	3.43	11.8	Scaf4_148046

[1]based on 1000 permutations